Thèse de Doctorat

Moustafa IBRAHIM

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'École Centrale de Nantes
sous le label de l'Université de Nantes Angers Le Mans

École doctorale : Sciences et Technologies de l'Information et Mathématiques

Discipline : Mathématiques et leurs interactions
Unité de recherche : Laboratoire de Mathématiques Jean Leray (LMJL)

Soutenue le 25 septembre 2014

Systèmes paraboliques dégénérés intervenant en mécanique des fluides et en médecine: analyse mathématique et numérique

JURY

Président :	**M. Iraj MORTAZAVI**, Professeur, Conservatoire National des Arts et Métiers
Rapporteurs :	**M. Laurent CHUPIN**, Professeur, Université Blaise Pascal **M. Robert EYMARD**, Professeur, Université Paris-Est Marne-la-Vallée
Examinateurs :	**M. Christophe BERTHON**, Professeur, Université de Nantes **M. Laurent DI MENZA**, Professeur, Université de Reims
Invité :	**M. Clément CANCÈS**, Maître de conférences, Université Paris VI
Directeur de thèse :	**M. Mazen SAAD**, Professeur, École Centrale de Nantes

Remerciements

Il m'a été donné de rencontrer une personne qui allait être pour moi une présence remarquable par sa richesse d'humanité. C'est avec une telle qualité de présence, toujours bienveillante et attentive, que Mazen Saad a donné réalité à mon désir d'explorer un passionnant domaine entre les mathématiques et la médecine. Il a toujours fait preuve d'une totale confiance, d'un soutien infiniment précieux et d'une gentillesse permanente à mon égard, qui m'ont permis de mener et d'achever mon travail de recherche pendant ces trois années. Je suis très heureux d'avoir eu la chance de bénéficier au niveau scientifique et humain de sa grande expérience, de sa vision mathématique, ainsi que de son immense culture. Trouver des mots assez forts pour dire "Merci" est une chose difficile, d'autant plus dans une langue qui n'est pas ma langue maternelle. Je souhaite ici remercier Mazen Saad pour son amitié et lui exprimer ma plus profonde gratitude.

Je souhaite remercier très vivement les professeurs qui se sont intéressés de près à mon travail : Laurent Chupin et Robert Eymard, qui ont accepté d'en être les rapporteurs et de donner de leur temps pour examiner cette thèse. Je remercie également Christophe Berthon, Laurent Di Menza et Iraj Mortazavi qui m'ont tous fait l'honneur et le plaisir de constituer le jury de ma thèse.

J'adresse également de vifs remerciements à Clément Cancès pour avoir accepter l'invitation et participer au jury de ma thèse, ainsi que pour sa contribution dans le quatrième chapitre de la thèse. Je le remercie pour son amitié et pour les nombreuses discussions que nous avons eues lors de sa visite à l'École Centrale de Nantes, au cours de laquelle il n'a pas hésité à partager avec moi son savoir et son expérience dans les domaines numérique et théorique. Je lui en suis très reconnaissant.

Je voudrais également exprimer mes plus chaleureux remerciements à Françoise Foucher, pour son aide et ses conseils pendant les trois années de la thèse. Je la remercie pour m'avoir donner l'occasion d'enseigner à l'École Centrale de Nantes, ainsi je n'oublie pas ses remarques et sa contribution à corriger les fautes d'orthographe dans l'introduction de la thèse.

Je remercie de façon générale les membres du laboratoire de mathématiques Jean Leray ainsi que les membres du département informatique et mathématiques de l'École Centrale de Nantes, dans lequel j'ai été accueilli avec bonne humeur et convivialité.

Au delà des horizons nantais, je tiens également à saluer et remercier toutes les personnes rencontrées dans les divers congrès ou séjours de recherche avec qui il a été agréable d'échanger ou de travailler. Je souhaite ici remercier Georges Chamoun pour son amitié et pour les beaux moments qu'on a passé ensemble lors de notre participation aux congrès à Sofia, Marseille et Taormina.

M'éloignant de la sphère des mathématiques, j'adresse ma reconnaissance et bien plus... à tous mes amis (en France et au Liban) pour leur soutien. Sans eux je n'aurais pas eu la joie de conti-

nuer sur ce long chemin studieux. Ce “succès” est également un peu le leur. Un grand Merci à Rita-Maria, Hanna, Rabih, Mohamad, Chady, Georges Salameh, Georges Massaad, Ali, Ibrahim, Michaël, Olivier... pour leur présence à ma soutenance et pour leur immense aide dans l'organisation du pot de la soutenance. Et je n'oublie pas à saluer Amine, Akram, Salim, Tala, Youssef, Hala,... pour leurs encouragements. À eux tous, je souhaite beaucoup de bien.

Je clos enfin ces remerciements en dédiant cette thèse de doctorat à ma famille : mon père, ma mère, mes frères (Ibrahim, Wissam, Abed, Majdi et Maher). Grâce à eux mon séjour en France et la poursuite de mes études jusqu'à ce stade ont été possibles. Je les remercie de tout mon cœur pour leur amour et leur soutien discret et essentiel. C'est avec leurs encouragements que j'ai pu accomplir ce travail de thèse. Il m'est impossible de trouver des mots pour dire à quel point je suis fier d'eux, et à quel point je les aime. J'en profite pour glisser un clin d'œil aussi au reste de la famille (au Liban, Syrie et Australie), sans qui je n'aurais pas eu le goût de me lancer dans cette aventure.

“At each given moment, there is only a fine layer between the “trivial” and the impossible. Mathematical discoveries are made in this layer.”

A.N. Kolmogorov

Résumé

Dans cette thèse, nous nous intéressons à l'analyse mathématique et numérique des systèmes paraboliques non linéaires dégénérés découlant, soit de la modélisation de la chimiotaxie, soit de la modélisation des fluides compressibles. Le modèle de chimiotaxie (Keller-Segel) proposé est un modèle de dynamique des populations décrivant l'évolution spatio-temporelle de la densité cellulaire et de la concentration chimiotactique. Pour ce modèle, nous étudions la formation de patterns en utilisant l'analyse de stabilité linéaire et le principe de Turing. Nous proposons ensuite un schéma numérique CVFE pour un modèle anisotrope de Keller-Segel. La construction de ce schéma est basée sur la méthode des éléments finis pour le terme de diffusion et sur la méthode des volumes finis classique pour le terme de convection. Nous montrons que ce schéma assure le principe de maximum discret et qu'il est consistent dans le cas où tous les coefficients de transmissibilité sont positifs. Par la suite, sur des maillages triangulaires généraux, nous proposons et analysons un schéma numérique CVFE non linéaire. Ce schéma est basé sur l'utilisation d'un flux numérique de Godunov pour le terme de diffusion, tandis que le terme de convection est approché au moyen d'un décentrage amont et d'un flux de Godunov. D'une part, le décentrage amont permet d'avoir le principe de maximum. D'autre part, le flux de Godunov assure que les solutions discrètes soient bornées sans restriction sur le maillage du domaine spatial ni sur les coefficients de transmissibilité. Nous réalisons différentes simulations numériques bi-dimensionnelles pour illustrer l'efficacité du schéma à tenir compte des hétérogénéités. Enfin, nous nous intéressons à une équation parabolique dégénérée contenant des termes dégénérés d'ordre 0 et 1 et décrivant un modèle de chimiotaxie-fluide ou l'écoulement d'un fluide compressible. Une formulation faible classique est souvent possible en absence des termes dégénérés d'ordre 0 et 1 ; tandis que dans le cas général, nous obtenons des solutions dans un sens affaibli vérifiant une formulation de type inégalité variationnelle. La définition des solutions faibles est adaptée à la nature de la dégénérescence des termes de dissipation.

Mots clés

Systèmes de réaction-diffusion, Formation de patterns, Chimiotaxie, Stabilité linéaire, Méthode de volumes finis, Méthode des éléments finis, Systèmes paraboliques dégénérés, Tenseurs anisotropes hétérogènes, Fluide compressible.

Abstract

In this thesis, we are interested in the mathematical and numerical analysis of nonlinear degenerate parabolic systems arising either from modeling the chemotaxis process, or from modeling compressible flows in porous media. The proposed chemotaxis model (Keller-Segel model) is a model of population dynamics describing the spatio-temporel evolution of the cell density and the chemical concentration. For this model, we study the pattern formation using the linear stability analysis as well as the principle of Turing. Then, we propose a numerical scheme (CVFE scheme) for an anisotropic Keller-Segel model. The construction of the scheme is based on the use of each of the finite element scheme for the diffusion term and the upwind finite volume scheme for the convective term. We show that the scheme is consistent and ensures the discrete maximum principle in the case where all the transmissibility coefficients are nonnegative. Thereafter, over general triangular meshes, we propose and analyze a nonlinear CVFE scheme. This scheme is based on the use of the Godunov flux function for the diffusion term, while the convective term is approximated by parts using an upwind finite volume scheme and a Godunov flux function. First, the upwind finite volume scheme allows of having the discrete maximum principle. On the other hand, the Godunov scheme ensures the boundedness of the discrete solutions without restrictions on the mesh nor on the transmissibility coefficients. Using this scheme, we realize some numerical simulations to illustrate the effectiveness of the scheme. Finally, we are interested in a degenerate parabolic equation containing degenerate terms of order 0 and 1 and describing a chemotaxis-fluid model or a displacement of compressible flows. Classical weak formulation is often possible in the absence of degenerate terms of order 0 and 1 ; while in the general case, we obtain weak solutions in the sense of verifying a weighted formulation. The definition of weak solutions is adapted to the nature of the degeneracy of the dissipative terms.

Keywords

Reaction-diffusion systems, Pattern formation, Chemotaxis, Linear stability, Finite volume method, Finite element method, Degenerate parabolic systems, Heterogeneous anisotropic tensors, Compressible fluid.

Table des matières

1

Introduction

Sommaire

1.1 Contexte général et scientifique

Les êtres humains ont longtemps eu une fascination par les motifs des pelages d'animaux (animal coat pattern en anglais). D'où provient la diversité des motifs des pelages d'animaux ? Pourquoi les léopards possèdent-ils des taches et ne possèdent pas des rayures ?

Toutes ces questions qui conduisent à comprendre le développement des motifs des pelages d'animaux chez les mammifères ont été un domaine d'études pour les biologistes et les chimistes. Une théorie sur la façon où ces motifs se produisent a été introduite par le mathématicien Alan Mathison Turing. Ce dernier a introduit un modèle pour lequel les motifs spatiaux peuvent se produire comme un résultat d'instabilité de la diffusion des substances chimiques morphogénétiques présentes dans la peau des animaux durant le stade embryonnaire du développement.

FIGURE 1.1 – Motifs du léopard, de girafe, du zèbre et du poisson.

1.1.1 Motivation biologique

La forme biologique d'un prémotif est inscrite dans les cellules : lors de la croissance de celles-ci, des produits chimiques (les morphogènes) se sont produites en concentration variable, et jouent suivant leur concentration le rôle d'activeur ou d'inhibiteur des pigments.

Pour l'obtention de tels motifs de pelages d'animaux, nous nous intéressons à un type spécifique de cellules appelées mélanocytes ; en effet, ce sont les cellules responsables de la production d'un pigment appelé la mélanine gouvernant la couleur de la peau.

Les biologistes pensent que les mélanocytes produisent de la mélanine basée sur la présence de certains activeurs et inhibiteurs des pigments. Chaque motif des pelages d'animaux est considéré comme étant le produit de certains activeurs et inhibiteurs [59].

1.1.2 Motivation mathématique (Principe de Turing)

En 1952, le mathématicien Alan Turing s'est intéressé à la morphogenèse et il a proposé un modèle basé sur des équations de **réaction–diffusion** (système de Turing) [76, 61]. Pour établir son modèle, Turing s'est appuyé sur la fameuse expérience réalisée par le biochimiste Belousov (1950) [62]. Ce dernier a utilisé des substances chimiques, et il a observé que la solution obtenue n'at-

FIGURE 1.2 – La réaction de Bolousov produisant des patterns temporels oscillants.

teint pas normalement l'état d'équilibre. En revanche il a observé qu'il y avait une apparition des patterns temporels oscillants jusqu'à épuisement d'un des réactifs.

De sa part, Turing a établi que, étant donné deux ensembles de substances chimiques, l'un est appelé **activeur** et l'autre est appelé **inhibiteur** : si les deux ensembles interagissent et se diffusent dans un domaine fini et si en plus ces substances se propagent dans l'espace avec des différents taux de diffusions, alors nous pouvons générer des patterns spatiaux et stationnaires.
Pour bien comprendre l'idée séminale de Turing, voici l'image la plus frappante : supposons qu'un feu (l'activeur) éclate dans une forêt sèche. Tout au début, il n'y a probablement pas de pompiers (les inhibiteurs) à proximité, mais avec leurs hélicoptères, ils peuvent dépasser le front de l'incen-

die et pulvériser sur les arbres des produits chimiques résistants au feu : quand le feu atteint les arbres traités, il s'éteint, l'incendie est arrêté.

Si plusieurs feux se déclarent spontanément et de façon aléatoire au sein de la forêt, plusieurs fronts d'incendie (des vagues d'activation) se propageront ; chaque incendie sera maîtrisé par les pompiers dans leurs hélicoptères (vagues d'inhibiteurs), mais il aura brulé la forêt tout autour de son foyer initial. À la fin de la saison sèche, la forêt présentera des zones noires formées par les arbres brulés et des zones vertes constituées d'arbres intacts.

Nous pouvons schématiser l'idée séminale de Turing comme suit : quand la diffusion n'est pas un paramètre essentiel c-à-d en absence de la diffusion (dans un milieu fortement brassé par exemple), les deux morphogènes réagissent et atteignent un état d'équilibre uniforme (pas de pattern). Dans le cas contraire, c-à-d en présence de la diffusion et sous certaines conditions appelées *conditions de Turing* (les morphogènes diffusent avec des vitesses différentes par exemple), une petite perturbation spatiale transforme l'état d'équilibre en un état instable engendrant des motifs spatiaux hétérogènes que nous appelons *Patterns de Turing*. En d'autres termes, sous les conditions de Turing la diffusion peut être déstabilisatrice et une petite perturbation spatiale peut être "instable" et engendrer un motif : une telle instabilité est appelée "diffusionnelle" ou bien "instabilité de turing".

1.1.3 Validation de l'idée de Turing

Après l'apparition de l'idée de Turing, plusieurs scientifiques ont travaillé pour vérifier cette idée, citons par exemple James Dickson Murray et Daniel Thomas. Murray s'est intéressé au modèle de Turing pour deux raisons. La première est que le chimiste Daniel Thomas (Université de Compiègne 1975) a travaillé sur le modèle de Turing et a pu démontrer par une expérience faite sur des produits chimiques spécifiques que le système de Turing peut engendrer des patterns spatiaux et stationnaires (patterns de Turing). Cette expérience était la première expérience confirmant que l'idée de Turing peut être fiable.

La deuxième raison pour laquelle Murray était intéressé par le modèle de Turing est que le modèle peut générer des patterns (des taches) comme ceux qui apparaissent sur la peau des animaux [61]. Murray a confirmé mathématiquement l'idée de Turing, et il a trouvé que, si en faisant juste varier le paramètre échelle de mesure γ (scale parameter) alors le modèle peut engendrer des patterns spatiaux qui prennent, suivant la valeur de γ, la forme de rayures ou de taches.

D'après ce qui précède, nous avons établit le rôle de la formation de patterns pour comprendre la formation de motifs de pelages d'animaux qui sont un résultat d'un principe de "réaction–diffusion" justifié par les chimistes : "activation, inhibition, diffusion". Un des exemples les plus connus et les plus populaires sur les systèmes de "réaction–diffusion" est la chimiotaxie.

1.1.4 La chimiotaxie

Un grand nombre des insectes et des animaux s'appuient sur une sensation aigüe de l'odorat, soit pour transmettre l'information entre les membres de l'espèce soit pour la prédation. Les produits chimiques impliqués dans ce processus sont appelés *phéromones*. Par exemple, la femelle du ver à soie libère une phéromone comme un attractif sexuel pour le mâle, qui dispose d'une antenne remarquablement efficace pour mesurer la concentration de la phéromone émise. Par conséquent, le mâle se déplace dans la direction de la forte et croissante concentration de la phéromone. La forte sensation de l'odorat pour la plupart des poissons est très importante pour la communication et la prédation ; par exemple, le requin dispose d'une sensation qui lui permet de sentir une goutte de sang diluée dans les tonnes d'eau, ce qui lui facilite la recherche de proie. L'exploitation la plus importante de la libération de phéromones est la direction du mouvement dirigé. Ce mouvement

dirigé par les produits chimiques est appelé la chimiotaxie, qui contrairement à la diffusion, dirige le mouvement vers un gradient de concentration.

Ce n'est pas seulement dans l'écologie des animaux et des insectes que la chimiotaxie est importante [43, 24, 68, 53, 21]. Elle peut être aussi cruciale dans les processus biologiques où il y a de nombreux exemples. Par exemple, quand une infection bactérienne envahit le corps, elle peut être attaquée par un mouvement de cellules vers la source par la suite de la chimiotaxie. Des preuves convaincantes suggèrent que les cellules de leucocytes dans le sang se dirigent vers la région d'inflammation bactérienne, pour l'attaquer, en se déplaçant suivant un gradient chimique causée par l'infection.

La modélisation théorique et mathématique de la chimiotaxie remonte aux travaux pionniers de Paltack [69] en 1950. Le modèle de Keller–Segel [51] introduit en 1970 reste le modèle le plus populaire et le plus connu pour le contrôle chimique des mouvement cellulaires. L'article [47] de Horstmann fournit une étude mathématique détaillée sur le modèle de Keller–Segel. Récemment, Bendahmane et al. [8] étudient l'existence des solutions pour un modèle de Keller–Segel dégénéré en tenant compte de l'effet de remplissage du volume.

L'analyse mathématique montre que les solutions du modèle de Keller–Segel peuvent modéliser des nombreux patterns spatiaux. Pour cela, l'étude de la formation de patterns pour le modèle de chimiotaxie est l'une des motivations principales de cette thèse.
Dans la suite, nous donnons un plan détaillé de la thèse avec les principaux résultats obtenus.

1.2 Plan de la thèse

Les travaux effectués dans cette thèse correspondent à l'étude théorique de la formation de patterns pour un modèle de chimiotaxie avec effet de remplissage du volume (volume-filling chemotaxis model en anglais) ; en plus, ils correspondent au développement d'une méthode numérique robuste et consistante capable de capturer la génération des patterns spatiaux pour le modèle étudié.

Dans le but d'obtenir un schéma numérique convergent sans restriction sur le maillage du domaine spatial ni sur les coefficients de transmissibilité, un schéma numérique est développé pour un modèle généralisé de Keller–Segel où nous prenons en compte une diffusion anisotrope et hétérogène de la densité cellulaire et de la concentration chimiotactique. Enfin, nous étudions l'existence de solutions faibles pour une équation parabolique dégénérée non linéaire découlant, soit de la modélisation de la chimiotaxie-fluide, soit de la modélisation d'un fluide compressible.

1.2.1 Chapitre 2 : Formation de patterns pour un modèle de chimiotaxie

Dans ce chapitre, nous nous intéressons à la formation de patterns spatiaux pour un système de réaction–convection–diffusion issue de la modélisation de l'effet de remplissage du volume pour un modèle de chimiotaxie. Ce système est donné par le couple des équations suivantes :

$$\begin{cases} \partial_t u - \operatorname{div}\left(d_1\left(q\left(u\right) - q'\left(u\right)u\right)\nabla u\right) + \operatorname{div}\left(uq\left(u\right)\chi\left(v\right)\nabla v\right) = f\left(u,v\right), \\ \partial_t v - d_2\Delta v = g\left(u,v\right). \end{cases} \tag{1.1}$$

Les deux équations du système (1.1) décrivent respectivement l'évolution en temps et en espace de la densité cellulaire u et de la concentration des substances chimiques v.
Le domaine occupé par les cellules et les substances chimiques est supposé fixe au cours du temps et est noté Ω, qui est un ensemble ouvert et borné de $\mathbb{R}^2$. Nous posons $Q_{t_\mathrm{f}} = \Omega \times (0, t_\mathrm{f})$, $\Sigma_{t_\mathrm{f}} = \partial\Omega \times (0, t_\mathrm{f})$. Le système (1.1) est complété par la donnée de conditions initiales

$$u\left(\mathbf{x},0\right) = u_0\left(\mathbf{x}\right) \geq 0, \qquad v\left(\mathbf{x},0\right) = v_0\left(\mathbf{x}\right) \geq 0, \tag{1.2}$$

pour tout $\mathbf{x} \in \Omega$ et par la donnée de conditions aux limites de type Neumann homogène

$$\left(d_1 \left(q\left(u\right) - q'\left(u\right) u\right) \nabla u - uq\left(u\right) \chi\left(v\right) \nabla v\right) \cdot \mathbf{n} = 0, \quad \nabla v \cdot \mathbf{n} = 0, \tag{1.3}$$

pour tout $(\mathbf{x}, t) \in \Sigma_{t_\mathrm{f}}$ où le vecteur $\mathbf{n}$ désigne la normale à $\partial\Omega$ sortante de Ω. L'existence d'une solution globale pour le modèle de chimiotaxie (1.1) a été obtenue pour la première fois avec Painter et Hillen dans [44], où la prolifération des cellules n'était pas prise en compte. Plus tard, Wrzosek [79] a démontré l'existence d'une solution globale du système (1.1), pour n'importe quelle dimension de l'espace et avec une cinétique non nulle de la densité cellulaire f et une probabilité de pression q linéaire dans $[0, \overline{u}]$, où $\overline{u}$ désigne le nombre total de cellules qui peuvent être logées dans n'importe quel site. Cette probabilité de pression reflète le fait que les cellules sont des particules solides ; elle est donnée par

$$q\left(u\right) = \begin{cases} 1 - \dfrac{u}{\overline{u}}, & 0 \leq u \leq \overline{u}, \\ 0, & u \geq \overline{u}. \end{cases} \tag{1.4}$$

Récemment, Wang et Hillen [78] ont montré l'existence d'une solution globale du système (1.1) pour une probabilité de pression plus réaliste (reflétant la propriété d'élasticité des cellules) en prenant une fonction non linéaire pour q.
Nous donnons les hypothèses portant sur le système (1.1) afin d'avoir des solutions globales ; elles sont données par

(A1) d_1 et d_2 sont deux constantes positives.

(A2) χ est une fonction de classe $\mathcal{C}^2\left(\mathbb{R}, \mathbb{R}\right)$ et telle que $\chi\left(v\right) > 0$.

(A3) La probabilité de pression q est une fonction concave et décroissante de classe $\mathcal{C}^3\left([0, \overline{u}]\right)$ satisfaisant la condition suivante : il existe un nombre critique $\overline{u}$ tel que $q\left(0\right) = 1$, $q\left(\overline{u}\right) = 0$, $0 < q\left(u\right) < 1$ pour $u \in (0, \overline{u})$ et $q\left(u\right) = 0$ pour tout $u > \overline{u}$.

(A4) f est une fonction de classe $\mathcal{C}^2\left(\mathbb{R} \times \mathbb{R}\right)$ satisfaisant $f\left(0, v\right) \geq 0$ pour tout $v \geq 0$. En outre, il existe une constante $0 < u_c < \overline{u}$ telle que $f\left(u_c, v\right) = 0$ et $f\left(u, v\right) < 0$ pour tout $u > u_c$ et $v \geq 0$.

(A5) g est une fonction de classe $\mathcal{C}^2\left(\mathbb{R} \times \mathbb{R}\right)$ satisfaisant $g\left(u, 0\right) \geq 0$ pour tout $u \geq 0$. En outre, il existe une constante $\overline{v} > 0$ telle que $g\left(u, \overline{v}\right) < 0$ pour tout $0 \leq u \leq \overline{u}$.

Analyse de stabilité linéaire pour le système (1.1)

Afin d'étudier la possibilité de générer des patterns spatiaux pour le système (1.1), nous appliquons l'idée séminale de Turing. Nous cherchons les états stationnaires stables (u_s, v_s) du système (1.1) en absence de toute variation spatiale ; ce qui amène à déterminer les états stationnaire stables pour le système suivant

$$\begin{aligned} &\partial_t u = f\left(u, v\right), \quad \partial_t v = g\left(u, v\right) && \text{dans } Q_{t_\mathrm{f}}, \\ &f\left(u_s, v_s\right) = g\left(u_s, v_s\right) = 0, && \text{dans } Q_{t_\mathrm{f}}. \end{aligned} \tag{1.5}$$

Pour cette raison, nous effectuons une linéarisation autour de l'état stationnaire (u_s, v_s) (cette linéarisation est possible d'après les hypothèses de différentiabilité (A4)–(A5) des fonctions f et g) ; nous posons $\mathbf{w} = (u - u_s, v - v_s)$, ainsi l'équation (1.5) devient pour $|\mathbf{w}|$ assez petit,

$$\partial_t \mathbf{w} = A\mathbf{w}, \qquad A = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix}_{u_s, v_s},$$

où A est la matrice de stabilité (c-à-d la matrice jacobienne du système (1.5) calculée à l'état stationnaire (u_s, v_s)). Désormais, et sauf mention contraire, les dérivées partielles des fonctions f et g sont évaluées à l'état stationnaire (u_s, v_s).
Par conséquent, les conditions nécessaires et suffisantes pour que l'état stationnaire homogène (u_s, v_s) soit linéairement stable, sont données par

$$\text{tr}(A) = f_u + g_v < 0, \qquad \det(A) = f_u g_v - f_v g_u > 0. \tag{1.6}$$

La deuxième étape de l'idée de Turing, consiste à déterminer les conditions pour lesquelles les états stationnaires stables (c-à-d ceux qui vérifient les conditions (1.6)) deviennent instables pour une petite perturbation spatiale. Pour cela, nous examinons une petite perturbation autour de l'état stationnaire (u_s, v_s) sous la forme suivante :

$$u = u_s + \varepsilon \tilde{u}\left(\mathbf{x}, t\right), \qquad v = v_s + \varepsilon \tilde{v}\left(\mathbf{x}, t\right), \quad \varepsilon \ll 1. \tag{1.7}$$

En substituant la forme (1.7) dans le système (1.1), et en utilisant un développement asymptotique formel, nous obtenons le système suivant, après avoir assimilé les termes du premier ordre suivant ε et éliminé les tildes pour la simplicité

$$\begin{cases} \partial_t u = d_1 \left(q\left(u_s\right) - q'\left(u_s\right) u_s\right) \Delta u - u_s q\left(u_s\right) \chi\left(v_s\right) \Delta v + u f_u + v f_v, \\ \partial_t v = d_2 \Delta v + u g_u + v g_v. \end{cases} \tag{1.8}$$

Rappelons que les dérivées partielles des fonctions f et g figurant dans le système (1.8) sont des fonctions de u_s et v_s.
Une analyse sera étendue afin d'obtenir les conditions nécessaires concernant l'instabilité de l'état stationnaire (u_s, v_s) ; cette analyse s'appuie d'une part sur l'utilisation du principe de superposition des solutions dans le but de linéariser les termes spatiaux, et d'autre part, sur la détermination de la relation de dispersion associée au système (1.8) donné par l'équation $\lambda^2 + a\left(k^2\right)\lambda + b\left(k^2\right) = 0$, où $a\left(k^2\right) = \left(\xi d_1 + d_2\right) k^2 - \left(f_u + g_v\right)$ et $b\left(k^2\right) = \xi d_1 d_2 k^4 + \left(\psi g_u - d_2 f_u - \xi d_1 g_v\right) k^2 + f_u\, g_v - f_v\, g_u$.
ξ et ψ sont respectivement les coefficients de Δu et Δv de la première équation du système (1.8).
Sous ces notations, les conditions nécessaires pour la générations de patterns spatiaux (conditions de Turing) pour le système (1.1) sont données par

$$\begin{aligned} &f_u + g_v < 0, \quad f_u\, g_v - f_v\, g_u > 0, \quad \psi g_u - d_2 f_u - \xi d_1 g_v < 0, \\ &\left(\psi g_u - d_2 f_u - \xi d_1 g_v\right)^2 - 4\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right) > 0. \end{aligned} \tag{1.9}$$

Le système (1.1) dépend de plusieurs paramètres qui interviennent par exemple dans les fonctions cinétiques ou dans le terme de convection du système ; ces paramètres sont appelés paramètres de bifurcation. La deuxième étape de l'étude de la formation de patterns, repose sur l'analyse de bifurcation pour localiser les éventuelles valeurs particulières des paramètres de bifurcations pour lesquels le comportement du système passe d'un état qualitatif à un autre. En effet, l'analyse de stabilité linéaire effectuée précédemment nous permet d'en déduire que la bifurcation se produit lorsque

$$\psi g_u - d_2 f_u - \xi d_1 g_v = -2\sqrt{\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)}. \tag{1.10}$$

Nous appliquons cette égalité pour déterminer les valeurs critiques de chaque paramètre de bifurcation. Par exemple, notons par $\vartheta = \chi\left(v_s\right)$ et remplaçons ξ par sa valeur dans l'équation (1.10) après avoir fixé tous les autres paramètres du système (1.1), nous obtenons la valeur critique ϑ_c donnée par

$$\vartheta_c = \frac{2\sqrt{\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)} - d_2 f_u - \xi d_1 g_v}{g_u u_s q\left(u_s\right)}.$$

D'après les conditions (1.9), la génération de patterns aura lieu lorsque $\vartheta > \vartheta_c$; ainsi, nous déterminons la gamme de nombres d'ondes instables associée au paramètre de bifurcation ϑ. Elle est définie par $k_1^2 < k^2 < k_2^2$, où k_1^2 et k_2^2 sont les zéros de l'équation $b\left(k^2\right) = 0$ données par

$$k_1^2 = \frac{\rho - \sqrt{\rho^2 - 4\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)}}{2\xi d_1 d_2},\ k_2^2 = \frac{\rho + \sqrt{\rho^2 - 4\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)}}{2\xi d_1 d_2},$$

avec $\rho = d_2 f_u + \xi d_1 g_v + \psi g_u$.
Dans le but de valider les résultats théoriques, la deuxième partie de ce chapitre est dédiée à l'examen numérique de la formation de patterns sous des paramètres spécifiques vérifiant les conditions de Turing (1.9). Nous utilisons la méthode de volumes finis pour discrétiser les équations du système (1.1)–(1.3) afin d'obtenir un schéma stable pour la discrétisation du terme de convection et vérifiant le principe de maximum (voir par exemple [78]) sur la solution (u, v).

Schéma des volumes finis

Soit $\mathcal{T}$ un maillage polygonal régulier et **admissible** (à préciser plus loin) du domaine Ω, constitué d'une famille de sous-domaines convexes, polygonaux non vides K de Ω appelés volumes de contrôle. Notons par h la taille de $\mathcal{T}$ définie par $h = \max_{K\in\mathcal{T}} \mathrm{diam}(K)$, où $\mathrm{diam}(K)$ désigne la plus grande distance entre chaque paire de sommets du polygone convexe K.

Pour tout $K \in \mathcal{T}$, notons par $\mathbf{x}_K$ le centre de K, $\mathcal{N}(K)$ l'ensemble de voisins de K. De plus, pour tout $L \in \mathcal{N}(K)$, nous notons par d_{KL} la distance entre $\mathbf{x}_K$ et $\mathbf{x}_L$, par σ_{KL} l'interface entre K et son voisin L, et par η_{KL} la normale unitaire à σ_{KL} dirigée de K vers L. La figure 1.3 donne une illustration des volumes de contrôle ainsi que de la distance entre les centres.

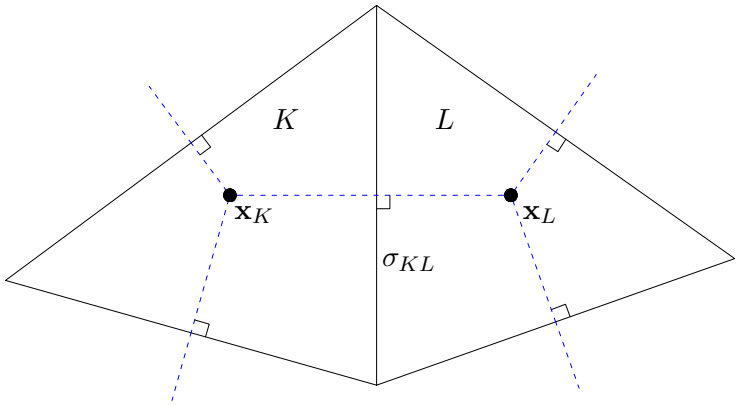


FIGURE 1.3 – Volumes de contrôle, centres, interfaces.

Pour tout $K \in \mathcal{T}$, nous désignons par $|K|$ la mesure de Lebesgue dans $\mathbb{R}^2$ du volume de contrôle K et par ∂K le bord de K formé par l'ensemble des arêtes de K.
$\mathcal{T}$ est un maillage orthogonal admissible dans le sens de Eymard–Gallouët–Herbin [31].

L'admissibilité de $\mathcal{T}$ implique que $\overline{\Omega} = \cup_{K\in\mathcal{T}} \overline{K}$, et pour tout volume de contrôle $L \in \mathcal{N}(K)$ (voisin du volume de contrôle K), le segment $[\mathbf{x}_K \mathbf{x}_L]$ joignant les centres de K et L est orthogonal à l'interface $\sigma_{KL} = \overline{K} \cap \overline{L}$.

Pour qu'un maillage conforme constitué de triangles vérifie la condition d'orthogonalité, il suffit de prendre le centre du volume de contrôle K, $\mathbf{x}_K$ comme étant le centre du cercle circonscrit au triangle K. En outre, pour assurer que le centre $\mathbf{x}_K$ soit dans le triangle K, nous imposons la condition que tous les angles du triangle K soient aigus.

L'avantage d'un tel maillage admissible, est de donner une approximation consistante de la dérivée dans la direction de la normale en utilisant uniquement deux points ; en effet par le développement de Taylor, nous déduisions que

$$\begin{aligned} &\mathbf{x}_L - \mathbf{x}_K = d_{KL}\eta_{KL}, \\ &\nabla u(\mathbf{x}) \cdot \eta_{KL} = \frac{u(\mathbf{x}_L) - u(\mathbf{x}_K)}{d_{KL}} + \mathcal{O}(h), \quad \forall \mathbf{x} \in \sigma = \overline{K} \cap \overline{L}. \end{aligned} \tag{1.11}$$

Une fois que le domaine est discrétisé, nous définissons l'espace discret $H_\mathcal{T}$ comme étant l'ensemble des fonctions constantes par morceaux sur les volumes de contrôle $K \in \mathcal{T}$. Chaque fonction $u_\mathcal{T} \in H_\mathcal{T}$ sera alors caractérisée par ses valeurs numériques $(u_K)_{K\in\mathcal{T}}$ telles que pour chaque volume de contrôle $K \in \mathcal{T}$, $u_{\mathcal{T}|_K} = u_K$. Plus précisément, la fonction discrète $u_\mathcal{T}$ s'écrit sous la forme d'une combinaison linéaire des fonctions caractéristiques $(\mathbf{1}_K)_{K\in\mathcal{T}}$ sous la forme suivante :

$$u_\mathcal{T}(\mathbf{x}) = \sum_{K\in\mathcal{T}} u_K \mathbf{1}_K(\mathbf{x}).$$

Nous donnons une définition du gradient discret $\nabla_\mathcal{T} u_\mathcal{T}$ qui est une fonction constante par diamant T_{KL}. Nous appelons un diamant T_{KL} associé à l'arête σ_{KL}, le polygone formé de quatre sommets $\mathbf{x}_K$, $\mathbf{x}_L$ et les deux sommets de l'arête σ_{KL} (voir figure 1.4). Le gradient discret est défini par

$$\nabla_\mathcal{T} u_{\mathcal{T}|_{T_\sigma}} = \begin{cases} 2\frac{u_L - u_K}{d_{KL}}\eta_{KL}, & \text{si } \sigma = \sigma_{KL}, \\ 0, & \text{si } \sigma \in \partial K \cap \partial\Omega. \end{cases}$$

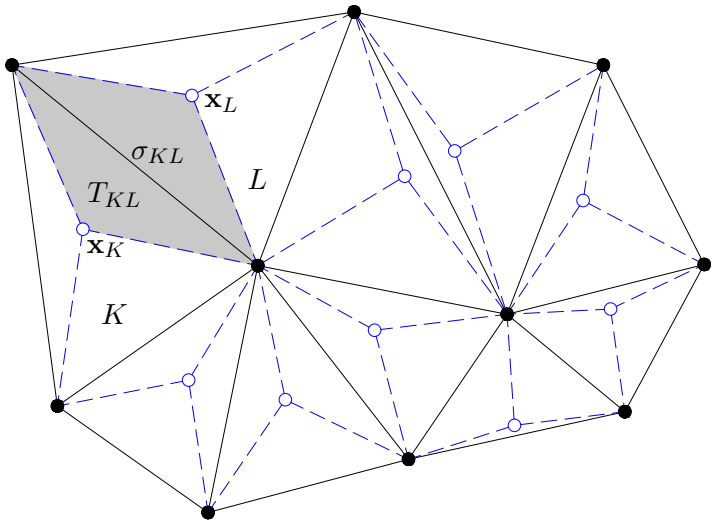


FIGURE 1.4 – Maillage admissible $\mathcal{T}$: volumes de contrôles, centres et diamants.

Ainsi, la norme dans $L^2(\Omega)$ de $\nabla_\mathcal{T} u_\mathcal{T}$ est donnée par

$$\|\nabla_\mathcal{T} u_\mathcal{T}\|^2_{L^2(\Omega)} = \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \frac{|\sigma_{KL}|}{d_{KL}} |u_K - u_L|^2.$$

La méthode des volumes finis consiste à intégrer directement les équations du modèle (contrairement à la méthode des éléments finis qui est basée sur les intégrales de la formulation variationnelle). Ainsi, les deux équations du système (1.1) sont intégrées sur chaque volume de contrôle $K \in \mathcal{T}$ et sur chaque intervalle de temps $]t_n, t_{n+1}[$, $n \in \mathbb{N}$ ($t_n = n\Delta t$, Δt est le pas de temps). Nous notons par A, Γ et Ψ les fonctions définies par $\Gamma(u) = q(u) - q'(u)\,u$, $A(u) = \int_0^u \Gamma(s)\,\mathrm{d}s$ et $\Psi(u) = uq(u)\,\chi$.

Pour les équations du système (1.1), et en appliquant le théorème de la divergence cela donne

$$\int_{t_n}^{t_{n+1}}\int_K \partial_t u \mathrm{d}\mathbf{x}\,\mathrm{d}t - d_1 \sum_{L\in\mathcal{N}(K)} \int_{t_n}^{t_{n+1}}\int_{\sigma_{KL}} \nabla A(u)\cdot\eta_{KL}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t + \sum_{L\in\mathcal{N}(K)} \int_{t_n}^{t_{n+1}}\int_{\sigma_{KL}} \Psi(u)\,\nabla v\cdot\eta_{KL}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_K f(u,v)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{1.12}$$

et

$$\int_{t_n}^{t_{n+1}}\int_K \partial_t v \mathrm{d}\mathbf{x}\,\mathrm{d}t - d_2 \sum_{L\in\mathcal{N}(K)} \int_{t_n}^{t_{n+1}}\int_{\sigma_{KL}} \nabla v\cdot\eta_{KL}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_K g(u,v)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \tag{1.13}$$

où, $\mathrm{d}\sigma(\mathbf{x})$ désigne la mesure de Lebesgue sur l'arête σ_{KL}.
Nous allons décrire brièvement une approximation de chaque terme des équations (1.12)–(1.13)

Les conditions initiales

$$u_K^0 = \frac{1}{|K|}\int_K u_0(\mathbf{x})\,\mathrm{d}\mathbf{x}, \qquad v_K^0 = \frac{1}{|K|}\int_K v_0(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Les termes d'évolution en temps

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_K \partial_t w \mathrm{d}\mathbf{x}\,\mathrm{d}t \approx \frac{1}{\Delta t}\int_K (w(\mathbf{x},t_{n+1}) - w(\mathbf{x},t_n))\,\mathrm{d}\mathbf{x} \approx |K|\,\frac{w_K^{n+1} - w_K^n}{\Delta t},\ w \equiv u \text{ ou } v.$$

Les termes de diffusion Le maillage admissible a été considéré afin de donner une approximation simple des flux de diffusion ; en effet, l'équation (1.11) nous permet d'écrire

$$\frac{d_1}{\Delta t}\sum_{L\in\mathcal{N}(K)} \int_{t_n}^{t_{n+1}}\int_{\sigma_{KL}} \nabla A(u)\cdot\eta_{KL}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx d_1 \sum_{L\in\mathcal{N}(K)} \frac{|\sigma_{KL}|}{d_{KL}}\left(A\left(u_L^{n+1}\right) - A\left(u_K^{n+1}\right)\right).$$

De la même manière, nous obtenons l'approximation du terme de diffusion de l'équation (1.13)

$$\frac{d_2}{\Delta t}\sum_{L\in\mathcal{N}(K)} \int_{t_n}^{t_{n+1}}\int_{\sigma_{KL}} \nabla v\cdot\eta_{KL}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx d_2 \sum_{L\in\mathcal{N}(K)} \frac{|\sigma_{KL}|}{d_{KL}}\left(v_L^{n+1} - v_K^{n+1}\right).$$

Le terme de convection L'approximation du terme de convection est différente de celle du terme de diffusion, ici nous voulons approcher le flux $\Psi(u)\,\nabla v\cdot\eta_{KL}$ (dépendant de u et de v) à l'interface σ_{KL} et à l'instant t_{n+1}. Le choix classique pour approcher un tel flux, consiste à utiliser un schéma décentré amont. Cette technique repose sur l'utilisation d'une fonction flux numérique G qui approche le flux $\Psi(u)\,\nabla v\cdot\eta_{KL}$ par le moyen des valeurs u_K, u_L et $\mathrm{d}V_{KL} := \frac{|\sigma_{KL}|}{d_{KL}}(v_L - v_K)$. Nous donnons les propriétés données sur la fonction G d'arguments $(a,b,c) \in \mathbb{R}^3$:

(i) Monotonie : $G(\cdot, b, c)$ est croissante pour tout $b, c \in \mathbb{R}$ et $G(a, \cdot, c)$ est décroissante pour tout $a, c \in \mathbb{R}$. Cette propriété est importante pour assurer le principe de maximum discret.

(ii) Conservativité : $G(a, b, c) = -G(b, a, -c)$ pour tout $a, b, c \in \mathbb{R}$. Cette condition nous permet d'établir l'intégration par parties discrète essentielle pour la convergence du schéma numérique.

(iii) Consistance : $G(a, a, c) = \Psi(a)\, c$ pour tout $a, c \in \mathbb{R}$.

(iv) Majoration locale : Il existe une constante $C > 0$ telle que $|G(a, b, c)| \leq C(|a| + |b|)|c|$, pour tout $a, b, c \in \mathbb{R}$.

(v) Continuité locale : Il existe un module de continuité $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ tel que $|G(a, b, c) - G(a', b', c)| \leq |c|\,\omega(|a - a'| + |b - b'|)$, pour tout $a, a', b, b', c \in \mathbb{R}$.

Nous prenons une fonction numérique G vérifiant les propriétés précédentes, par suite l'approximation du terme de convection est définie par

$$\frac{1}{\Delta t} \sum_{L \in \mathcal{N}(K)} \int_{t_n}^{t_{n+1}} \int_{\sigma_{KL}} \Psi(u)\, \nabla v \cdot \eta_{KL}\, \mathrm{d}\sigma(\mathbf{x})\, \mathrm{d}t \approx \sum_{L \in \mathcal{N}(K)} G\left(u_K^{n+1}, u_L^{n+1}; \mathrm{d}V_{KL}^{n+1}\right)$$

En résumé, le schéma de type volumes finis proposé pour la discrétisation du problème (1.1)–(1.3) consiste à chercher $U = \left(u_K^{n+1}\right)_{K \in \mathcal{T}, n \in [0,\ldots,N]}$ et $V = \left(v_K^{n+1}\right)_{K \in \mathcal{T}, n \in [0,\ldots,N]}$ solution de

$$u_K^0 = \frac{1}{|K|} \int_K u_0(\mathbf{x})\, \mathrm{d}\mathbf{x}, \qquad v_K^0 = \frac{1}{|K|} \int_K v_0(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{1.14}$$

et

$$\begin{aligned} |K| \frac{u_K^{n+1} - u_K^n}{\Delta t} - d_1 \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{d_{KL}} \left(A\left(u_L^{n+1}\right) - A\left(u_K^{n+1}\right)\right) \\ + \sum_{L \in N(K)} G\left(u_K^{n+1}, u_L^{n+1}; \mathrm{d}V_{KL}^{n+1}\right) = f\left(u_K^{n+1}, v_K^{n+1}\right), \\ |K| \frac{v_K^{n+1} - v_K^n}{\Delta t} - d_2 \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{d_{KL}} \left(v_L^{n+1} - v_K^{n+1}\right) = g\left(u_K^n, v_K^{n+1}\right). \end{aligned} \tag{1.15}$$

Sous les hypothèses (A1)–(A5) et supposons que $(u_0, v_0) \in \left(L^\infty(Q_{t_\mathrm{f}})\right)^2$ et tel que $0 \leq u_0 \leq 1$, $v_0 \geq 0$. Alors le schéma (1.14)–(1.15) converge dans $L^2(Q_{t_\mathrm{f}})$ quand $\Delta t, h \to 0$, pour une sous-suite, vers une solution faible (u, v) du problème (1.1)–(1.3).

La résolution de ce schéma nécessite toujours la résolution d'un système non linéaire. Pour cela, nous avons implémenté la méthode de Newton couplé avec l'algorithme du bigradient.

Enfin, des tests numériques en dimension deux sont faits pour simuler la formation de patterns pour le modèle (1.1)–(1.3) et pour montrer la robustesse du schéma (1.14)–(1.15) à capturer les patterns spatiaux.

1.2.2 Chapitre 3 : Un schéma volumes finis éléments finis pour capturer les patterns pour un modèle de chimiotaxie

Dans ce chapitre, nous nous sommes intéressés à un modèle de Keller–Segel similaire à celui de la section précédente, et pour lequel nous avons ajouté des tenseurs pour les termes de diffusion et

pour le terme de convection. Plus précisément, nous considérons le système suivant :

$$\begin{cases} \partial_t u - \operatorname{div}\left(\Lambda\left(\mathbf{x}\right) a\left(u\right) \nabla u - \Lambda\left(\mathbf{x}\right) \chi\left(u\right) \nabla v\right) = f\left(u\right) & \text{dans } Q_{t_{\mathrm{f}}} = \Omega \times \left(0, t_{\mathrm{f}}\right), \\ \partial_t v - \operatorname{div}\left(D\left(\mathbf{x}\right) \nabla v\right) = g\left(u, v\right) & \text{dans } Q_{t_{\mathrm{f}}} = \Omega \times \left(0, t_{\mathrm{f}}\right), \end{cases} \tag{1.16}$$

avec les conditions aux limites données sur le bord $\Sigma_T := \partial\Omega \times \left(0, t_{\mathrm{f}}\right)$ par :

$$\left(\Lambda\left(\mathbf{x}\right) a\left(u\right) \nabla u - \Lambda\left(\mathbf{x}\right) \chi\left(u\right) \nabla v\right) \cdot \mathbf{n} = 0, \qquad D\left(\mathbf{x}\right) \nabla v \cdot \mathbf{n} = 0, \tag{1.17}$$

où $\mathbf{n}$ est le vecteur normal à $\partial\Omega$ sortant de Ω. Les conditions initiales sont données par :

$$u\left(\mathbf{x}, 0\right) = u_0\left(\mathbf{x}\right), \quad v\left(\mathbf{x}, 0\right) = v_0\left(\mathbf{x}\right), \quad \mathbf{x} \in \Omega. \tag{1.18}$$

Le domaine occupé par les cellules et les substances chimiques est toujours noté par Ω, un ensemble ouvert et borné de $\mathbb{R}^2$. Nous donnons également les hypothèses portant sur la diffusion de la densité cellulaire a, la sensibilité chimiotactique χ, les tenseurs de diffusion Λ et D, les cinétiques et les conditions initiales :

(A1) $a : [0,1] \longrightarrow \mathbb{R}$ est une fonction continue telle que : $a\left(0\right) \geq 0$, $a\left(1\right) \geq 0$ et $a(u) > 0$ pour tout $0 < u < 1$.

(A2) $\chi : [0,1] \longrightarrow \mathbb{R}$ est une fonction differentiable telle que : $\chi(0) = \chi(1) = 0$ et $\chi(u) > 0$ pour tout $0 < u < 1$.

(A3) Les tenseurs de diffusion Λ et D sont deux tenseurs symétriques bornés, uniformément positifs sur Ω, c-à-d tel que :
$\forall \mathbf{w} \neq 0$, il existe deux constantes strictement positives T_- et T_+ telles que $0 < T_- \left|\mathbf{w}\right|^2 \leq \langle T(\mathbf{x})\mathbf{w}, \mathbf{w}\rangle \leq T_+ \left|\mathbf{w}\right|^2 < \infty$, $T = \Lambda$ ou D.

(A4) La fonction $f : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction continue telle que : $f\left(0\right) \geq 0$ et $f\left(1\right) \leq 0$.

(A5) Les conditions initiales u_0 et v_0 sont deux fonctions dans $L^{\infty}\left(\Omega\right)$ tel que : $0 \leq u_0 \leq 1$ et $v_0 \geq 0$.

Notons par $A : \mathbb{R} \longrightarrow \mathbb{R}$ la fonction Lipschitzienne et croissante définie par $A\left(u\right) = \int_0^u a\left(s\right)\, \mathrm{d}s$, pour tout $u \in \mathbb{R}$.

Definition 1.1 (Solution faible)**.** Sous les hypothèses (A1)–(A5). Le couple (u, v) est dit solution faible du système (1.16)–(1.17) si il vérifie :

$$\begin{aligned} &0 \leq u\left(\mathbf{x}, t\right) \leq 1, \ v\left(\mathbf{x}, t\right) \geq 0 \text{ pour presque tout } \left(\mathbf{x}, t\right) \in Q_{t_{\mathrm{f}}}, \\ &A\left(u\right) \in L^2\left(0, t_{\mathrm{f}}; H^1\left(\Omega\right)\right), \\ &v \in L^{\infty}\left(Q_{t_{\mathrm{f}}}\right) \cap L^2\left(0, t_{\mathrm{f}}; H^1\left(\Omega\right)\right), \end{aligned}$$

et pour tout $\varphi, \psi \in \mathcal{D}\left(\overline{\Omega} \times \left[0, t_{\mathrm{f}}\right)\right)$

$$\begin{aligned} -\int_{\Omega} u_0\left(\mathbf{x}\right) \varphi\left(\mathbf{x}, 0\right) \mathrm{d}\mathbf{x} - \iint_{Q_{t_{\mathrm{f}}}} u \partial_t \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t + \iint_{Q_{t_{\mathrm{f}}}} \Lambda\left(\mathbf{x}\right) \nabla A\left(u\right) \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ - \iint_{Q_{t_{\mathrm{f}}}} \Lambda\left(\mathbf{x}\right) \chi\left(u\right) \nabla v \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t = \iint_{Q_{t_{\mathrm{f}}}} f\left(u\right) \varphi\left(\mathbf{x}, t\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t, \\ -\int_{\Omega} v_0\left(\mathbf{x}\right) \psi\left(\mathbf{x}, 0\right) \mathrm{d}\mathbf{x} - \iint_{Q_{t_{\mathrm{f}}}} v \partial_t \psi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ + \iint_{Q_{t_{\mathrm{f}}}} D\left(\mathbf{x}\right) \nabla v \cdot \nabla \psi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t = \iint_{Q_{t_{\mathrm{f}}}} g\left(u, v\right) \psi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t. \end{aligned}$$

L'intention de ce chapitre est de construire un schéma numérique robuste et convergent pour l'approximation du système (1.16)–(1.18). La difficulté repose sur la présence d'un tenseur anisotrope dans les termes de diffusion ; en effet, la méthode des volumes finis classique, malgré sa capacité à assurer la stabilité pour le terme de convection, ne permet pas la manipulation des problèmes de diffusion avec des tenseurs anisotropes, même si la condition d'orthogonalité est satisfaite. La raison de la non efficacité de la méthode de volumes finis, est que contrairement à l'approximation (1.11), il n'y a pas un moyen simple et direct pour approcher les flux de diffusion avec des tenseurs anisotropes et hétérogènes. Néanmoins, il est bien connu que la méthode des éléments finis nous permet de réaliser une discrétisation très simple des termes de diffusion avec des tenseurs pleins et sans imposer aucune restriction sur la géométrie du maillage ; par contre, des instabilités numériques peuvent se produire dans le cas d'une convection dominée. Une idée assez intuitive est donc de combiner les deux méthodes en donnant une discrétisation par la méthode des éléments finis conformes pour les termes de diffusion et une discrétisation par la méthode de volumes finis pour les autres termes. Par conséquent, nous construisons et nous étudions l'analyse de convergence d'un nouveau schéma appelé schéma CVFE afin de simuler et capturer la formation de patterns pour le modèle de chimiotaxie anisotrope.

Discrétisation du problème continu avec le schéma CVFE

Nous nous plaçons dans le cas de figure où les bordures du domaine Ω sont fixes au cours du temps. La construction de la solution approchée nécessite l'introduction de deux différentes discrétisations spatiales du domaine Ω, à savoir la *discrétisation primale* et la *discrétisation duale*. La discrétisation primale $\mathcal{T}$ est une triangulation conforme du domaine Ω constituée d'un nombre fini de triangles qui forment une partition du domaine Ω c-à-d $\overline{\Omega} = \bigcup_{K\in\mathcal{T}} \overline{K}$ et $K \cap K' = \emptyset$ si $K \neq K'$. Pour chaque triangle $K \in \mathcal{T}$, nous notons par $\mathbf{x}_K$ son centre de gravité, par h_K son diamètre, par ρ_K le diamètre du cercle circonscrit au triangle K et par $|K|$ la mesure de Lebesgue du triangle K. Ainsi, pour chaque sommet S du triangle K, nous notons par $\mathcal{K}_S$ l'ensemble de triangles admettant S comme sommet et par $\mathcal{E}_S$ l'ensemble des arêtes admettant S comme extrémité.

L'objectif de la technique CVFE est de reconstruire la solution approchée aux sommets du maillage primal. Pour ce faire, une nouvelle partition du domaine est définie de telle sorte que chaque sommet du maillage primal ne soit pas inclus que dans un triangle de la nouvelle partition. Le maillage issu de ce deuxième partitionnement est appelé *maillage dual* noté par $\mathcal{M}$ et les polygones qui le composent *cellules duales* ou *volumes de contrôle duaux*. Pour chaque sommet S du triangle $K \in \mathcal{T}$, il existe un unique volume de contrôle dual M construit autour du sommet S en joignant les centres de gravité $\mathbf{x}_K$ des triangles $K \in \mathcal{K}_S$ avec les milieux des arêtes $\sigma \in \mathcal{E}_S$. Le centre de M est noté par $\mathbf{x}_M$, il est confondu avec le sommet S autour duquel le volume de contrôle dual M est construit. Finalement, pour chaque paire de volumes de contrôle duaux M et M', nous notons par $\sigma^K_{M,M'}$ le segment appartenant au triangle K et joignant le centre de gravité $\mathbf{x}_K$ avec le milieu de l'arête $[\mathbf{x}_M\mathbf{x}_{M'}]$ où $K \in \mathcal{T}$ est un triangle tel que $M \cap K \neq \emptyset$ et $M' \cap K \neq \emptyset$. La figure 1.5 donne une illustration du maillage primal et du maillage dual correspondant.

Dans tout ce chapitre, nous limitons notre étude au cas d'une discrétisation uniforme en temps avec un pas de temps donné par $\Delta t = t_{\mathrm{f}}/\left(N+1\right)$, $N \in \mathbb{N}^*$ et par suite $t_n = n\Delta t$, $0 \leq n \leq N+1$; ainsi, nous notons par $h = \max_{M\in\mathscr{M}} \mathrm{diam}\left(M\right)$ le pas d'espace (la taille du maillage).

Pour chacune des deux reconstructions spatiales précédentes, nous définissons deux solutions approchées au sens du schéma CVFE :

(i) Une solution volume fini $(u_{\mathcal{M},\Delta t}, v_{\mathcal{M},\Delta t})$ est définie comme étant une fonction constante

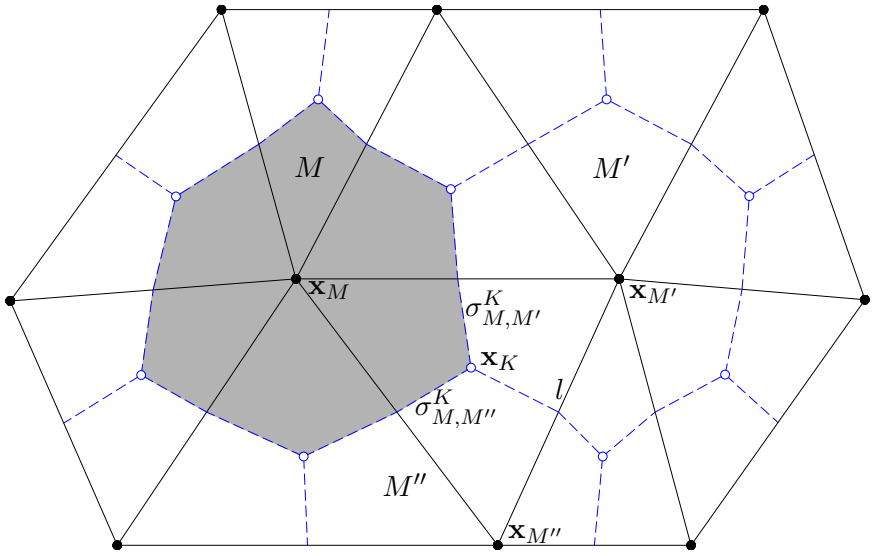


FIGURE 1.5 – Le maillage triangulaire primal $\mathcal{T}$ et le maillage dual correspondant $\mathcal{M}$: volumes de contrôle, centres et interfaces.

par morceaux sur le maillage dual et telle que

$$\begin{aligned}
&\left(u_{\mathcal{M},\Delta t}\left(\mathbf{x},0\right), v_{\mathcal{M},\Delta t}\left(\mathbf{x},0\right)\right) = \left(u_M^0, v_M^0\right) && \forall \mathbf{x} \in \overset{\circ}{M},\ M \in \mathcal{M},\\
&\left(u_{\mathcal{M},\Delta t}\left(\mathbf{x},t\right), v_{\mathcal{M},\Delta t}\left(\mathbf{x},t\right)\right) = \left(u_M^{n+1}, v_M^{n+1}\right) && \forall \mathbf{x} \in \overset{\circ}{M},\ M \in \mathcal{M}, \forall t \in (t_n, t_{n+1}],
\end{aligned}$$

où u_M^0 (resp. v_M^0) représente la valeur moyenne de la fonction u_0 (resp. v_0) sur M. L'espace discret de ces fonctions est noté $\mathcal{X}_{\mathcal{M},\Delta t}$.

(ii) Une solution élément fini $v_{\mathcal{T},\Delta t}$ est définie comme étant une fonction continue et affine par triangle et telle que

$$\begin{aligned}
&v_{\mathcal{T},\Delta t}\left(\mathbf{x},0\right) = v_{\mathcal{T}}^0\left(\mathbf{x}\right) && \forall \mathbf{x} \in \Omega,\\
&v_{\mathcal{T},\Delta t}\left(\mathbf{x},t\right) = v_{\mathcal{T}}^{n+1}\left(\mathbf{x}\right) && \forall \mathbf{x} \in \Omega, \forall t \in (t_n, t_{n+1}],
\end{aligned}$$

où $v_{\mathcal{T}}^{n+1}\left(\mathbf{x}\right) := \sum_{M\in\mathcal{M}} v_M^{n+1}\varphi_M\left(\mathbf{x}\right)$ et $v_{\mathcal{T}}^0\left(\mathbf{x}\right) := \sum_{M\in\mathcal{M}} v_M^0\varphi_M\left(\mathbf{x}\right)$. L'espace discret de ces fonctions est noté $\mathcal{H}_{\mathcal{T},\Delta t}$. $(\varphi_M)_{M\in\mathcal{M}}$ est la base canonique de $\mathcal{H}_{\mathcal{T},\Delta t}$.

La fonction A est non linéaire, nous notons $A_{\mathcal{T},\Delta t} = A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}\right)$ la reconstruction éléments finis correspondante dans $\mathcal{H}_{\mathcal{T},\Delta t}$, et $A_{\mathcal{M},\Delta t} = A\left(u_{\mathcal{M},\Delta t}\right)$ la reconstruction volumes finis correspondante dans $\mathcal{X}_{\mathcal{M},\Delta t}$. Pour les équations du système (1.16), et en appliquant la formule de Green–Gauss cela donne

$$\begin{aligned}
&\int_{t_n}^{t_{n+1}}\int_M \partial_t u \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla A(u)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t\\
&\qquad + \sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \chi\left(u\right)\Lambda\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_M f\left(u\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t, \qquad (1.19)\\
&\int_{t_n}^{t_{n+1}}\int_M \partial_t v \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma D\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_M g\left(u,v\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t,
\end{aligned}$$

où, $\eta_{M,\sigma}$ est le vecteur normal unitaire à σ sortant de M et $\mathrm{d}\sigma(\mathbf{x})$désigne la mesure de Lebesgue sur l'arête σ.
Nous allons décrire brièvement l'approximation de chaque terme des équations du système (1.19).

Les conditions initiales

$$u_M^0 = \frac{1}{|M|}\int_M u_0(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad v_M^0 = \frac{1}{|M|}\int_M v_0(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Les termes d'évolution en temps

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_M \partial_t w \mathrm{d}\mathbf{x}\,\mathrm{d}t \approx \frac{1}{\Delta t}\int_M \left(w\left(\mathbf{x},t_{n+1}\right)-w\left(\mathbf{x},t_n\right)\right)\mathrm{d}\mathbf{x} \approx |M|\,\frac{w_M^{n+1}-w_M^n}{\Delta t},\ w\equiv u \text{ ou } v.$$

Les termes de diffusion

$$\frac{1}{\Delta t}\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla A\left(u\right)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx \sum_{\sigma\subset\partial M}\int_\sigma \Lambda\nabla A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}\left(\mathbf{x},t_{n+1}\right)\right)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x}).$$

Nous utilisons la reconstruction éléments finis $A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}\right)$ ainsi que la définition des fonctions de base $(\varphi_M)_{M\in\mathcal{M}}$ pour déduire l'approximation suivante

$$\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla A(u)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\mathrm{d}t \approx \Delta t\sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset}\Lambda_{M,M'}^K\big(A\left(u_{M'}^{n+1}\right)-A\big(u_M^{n+1}\big)\big).$$

Le coefficient $\Lambda_{M,M'}^K$ est appelé le coefficient de transmissibilité défini par

$$\Lambda_{M,M'}^K = -\int_K \Lambda\left(\mathbf{x}\right)\nabla\varphi_M\left(\mathbf{x}\right)\cdot\nabla\varphi_{M'}\left(\mathbf{x}\right)\mathrm{d}\mathbf{x}.$$

De la même manière, nous obtenons une approximation du terme de diffusion de l'équation en v

$$\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx \Delta t\sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset} D_{M,M'}^K\big(v_{M'}^{n+1}-v_M^{n+1}\big),$$

avec $D_{M,M'}^K = -\displaystyle\int_K D\left(\mathbf{x}\right)\nabla\varphi_M\left(\mathbf{x}\right)\cdot\nabla\varphi_{M'}\left(\mathbf{x}\right)\mathrm{d}\mathbf{x}.$

Le terme de convection L'approximation du terme de convection est similaire à celle de la section précédente, nous voulons approcher le flux $\Lambda\left(\mathbf{x}\right)\chi\left(u_{\mathcal{M},\Delta t}\right)\nabla v_{\mathcal{T},\Delta t}\cdot\eta_{M,\sigma}$ à l'interface $\sigma\subset\partial M\cap\overline{K}$ et à l'instant t_{n+1}. Nous utilisons une fonction flux numérique G vérifiant les conditions (i)–(v) (données dans la section précédente) et qui approche le flux $\Lambda\left(\mathbf{x}\right)\chi\left(u_{\mathcal{M},\Delta t}\right)\nabla v_{\mathcal{T},\Delta t}\cdot\eta_{M,\sigma}$ par l'intermédiaire de u_M, $u_{M'}$ et l'approximation du gradient de v sur l'interface σ. En conclusion, l'approximation du terme de convection est donnée par

$$\Delta t\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \chi\left(u\right)\Lambda\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx$$
$$\sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset}\left|\sigma_{M,M'}^K\right| G\big(u_M^{n+1},u_{M'}^{n+1};\,\mathrm{d}V_{M,M'}^K\big).$$

En résumé, le schéma CVFE proposé pour la discrétisation du problème (1.16)–(1.18) consiste à chercher $U = \left(u_M^{n+1}\right)_{M\in\mathcal{M},n\in[0,\dots,N]}$ et $V = \left(v_M^{n+1}\right)_{M\in\mathcal{M},n\in[0,\dots,N]}$ solution de

$$u_M^0 = \frac{1}{|M|}\int_M u_0\left(\mathbf{x}\right)\mathrm{d}\mathbf{x}, \qquad v_M^0 = \frac{1}{|M|}\int_M v_0\left(\mathbf{x}\right)\mathrm{d}\mathbf{x}, \tag{1.20}$$

et

$$\begin{aligned} |M|\,\frac{u_M^{n+1}-u_M^n}{\Delta t} - \sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset} \Lambda_{M,M'}^K\left(A\left(u_{M'}^{n+1}\right)-A\left(u_M^{n+1}\right)\right) \\ + \sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset} \left|\sigma_{M,M'}^K\right| G\left(u_M^{n+1},u_{M'}^{n+1};\,\mathrm{d}V_{M,M'}^K\right) = |M|\,f\left(u_M^{n+1}\right), \\ |M|\,\frac{v_M^{n+1}-v_M^n}{\Delta t} - \sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset} D_{M,M'}^K\left(v_{M'}^{n+1}-v_M^{n+1}\right) = |M|\,g\left(u_M^n,v_M^{n+1}\right). \end{aligned} \tag{1.21}$$

Dans tout ce qui suit, nous considérons les hypothèses suivantes

(P1) $\Lambda_{M,M'}^K \geq 0$ et $D_{M,M'}^K \geq 0$, pout tout $M, M' \in \mathcal{M}$, et pour tout $K \in \mathcal{T}$.

(P2) il existe une constante $\kappa_{\mathcal{T}}$ strictement positive telle que : $\min_{K\in\mathcal{T}} \frac{|K|}{\mathrm{diam}(K)^2} \geq \kappa_T$.

Estimations discrètes, existence et convergence du schéma

Nous donnons des estimations a priori sur les solutions du schéma non linéaire (1.20)–(1.21) nécessaires pour établir l'existence d'une solution du schéma. Ensuite, nous présentons les résultats de compacité sur les solutions du schéma (translations en temps et en espace). Finalement, nous donnons le résultat principal de ce chapitre concernant la convergence du schéma basé sur l'utilisation du théorème de compacité de Kolmogorov.

Principe de maximum discret

Proposition 1.2. *Sous l'hypothèse* (P1)*, supposons que* $(u_M^n, v_M^n)_{M\in\mathcal{M},\, n\in\{0,\dots,N+1\}}$ *est une solution du schéma non linéaire* (1.20)–(1.21)*. Alors, nous avons* $0 \leq u_M^n \leq 1$ *et* $0 \leq v_M^n$ *pour tout* $M \in \mathcal{M}$*, et pour tout* $n \in \{0,\dots,N+1\}$*. En outre, il existe une constante positive* $\rho = \|v_0\|_\infty + \alpha t_{\mathrm{f}}$*, telle que* $v_M^n \leq \rho$*, pour tout* $n \in \{0,\dots,N+1\}$*.*

Nous donnons l'idée de la preuve pour u_M^{n+1} (la preuve pour v_M^{n+1} se traite de la même façon). Nous prenons un volume de contrôle dual M tel que $u_M^{n+1} = \min\left\{u_{M'}^{n+1}\right\}_{M'\in\mathcal{M}}$ et multiplions la première équation du système (1.21) par $-\left(u_M^{n+1}\right)^-$ et faisons la sommation sur $M \in \mathcal{M}$. Nous obtenons la positivité de u_M^{n+1} en utilisant la formule d'intégration par parties discrètes, l'hypothèse sur la positivité des coefficients de transmissibilité et par le prolongement continue par zéro des fonctions $\chi\left(u\right)$ et $f\left(u\right)$ pour $u \leq 0$. D'une manière similaire, et en multipliant la première équation du système (1.21) par $\left(u_M^{n+1}-1\right)^+$, nous obtenons que $u_M^{n+1} \leq 1$.

Estimations a priori Maintenant, nous donnons des propriétés discrètes sur le schéma non linéaire (1.20)–(1.21).

Proposition 1.3. *Sous les hypothèses* (P1) *et* (P2)*, supposons que* $(u_M^n, v_M^n)_{M\in\mathcal{M},\, n\in\{0,\dots,N+1\}}$ *est une solution du schéma non linéaire* (1.20)–(1.21)*. Alors, il existe une constante* $C > 0$ *(indépendante de* h *et* Δt*) et telle que*

$$\sum_{M\in\mathcal{M}} |M| \left(v_M^{n+1}\right)^2 + \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} D_{M,M'}^K \left(v_M^{n+1} - v_{M'}^{n+1}\right)^2 \leq \sum_{M\in\mathcal{M}} |M| \left(v_M^0\right)^2 \leq C,$$

pour tout $n \in \{0, \dots, N+1\}$.

Notons par $\mathcal{B} : \mathbb{R} \longrightarrow \mathbb{R}$ la fonction définie par $\mathcal{B}(u) = \int_0^u A(s)\, \mathrm{d}s$. Nous avons une estimation similaire à celle de la Proposition 1.3, elle est donnée par la proposition suivante

Proposition 1.4. *Pour tout* $n \in \{0, \dots, N+1\}$*, il existe une constante* $C > 0$ *(indépendante de* h *et* Δt*) et telle que*

$$\sum_{M\in\mathcal{M}} |M| \mathcal{B}\left(u_M^{n+1}\right) + \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} \Lambda_{M,M'}^K \left(A(u_M^{n+1}) - A(u_{M'}^{n+1})\right)^2 \leq C,$$

En utilisant les propositions 1.3 et 1.4, le théorème du point fixe de Brouwer assure l'existence d'au moins une solution $(u_M^n, v_M^n)_{M\in\mathcal{M},\, n\in\{0,\dots,N+1\}}$ du problème discret non linéaire (1.20)–(1.21).

Estimations de compacité sur la solution discrète Nous donnons ici les translations en temps et en espace nécessaires pour montrer la convergence du schéma (1.20)–(1.21). En particulier, nous avons la proposition suivante :

Proposition 1.5. *Sous les hypothèses* (P1) *et* (P2)*, il existe une constante* $C > 0$ *indépendante de* h *et* τ *et telle que*

$$\iint_{\Omega\times(0,t_{\mathrm{f}}-\tau)} |w_{\mathcal{M}_h,\Delta t}(t+\tau, \mathbf{x}) - w_{\mathcal{M}_h,\Delta t}(\mathbf{x}, t)|^2 \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C(\tau + \Delta t), \quad \textit{pour tout } \tau \in (0, t_{\mathrm{f}}),$$

et

$$\int_0^{t_{\mathrm{f}}} \int_{\Omega'} |w_{\mathcal{M}_h,\Delta t}(\mathbf{x}+\mathbf{y}, t) - w_{\mathcal{M}_h,\Delta t}(\mathbf{x}, t)| \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C(|\mathbf{y}| + h), \quad \textit{pour tout } \mathbf{y} \in \mathbb{R}^2,$$

avec $\Omega' = \{\mathbf{x} \in \Omega, [\mathbf{x}, \mathbf{x}+\mathbf{y}] \subset \Omega\}$ *et* $w_{\mathcal{M}_h,\Delta t} = A(u_{\mathcal{M}_h,\Delta t})$ *ou* $v_{\mathcal{M}_h,\Delta t}$.

Avant de passer au résultat de convergence, nous montrons un lemme essentiel qui assure que la reconstruction éléments finis et la reconstruction volumes finis se rapprochent l'une de l'autre quand la taille du maillage h tend vers zéro. Ce résultat est donné par le lemme suivant :

Lemma 1.6. *Les suites* $(A(u_{\mathcal{M}_h,\Delta t}) - A_{\mathcal{T}_h}(u_{\mathcal{T}_h,\Delta t}))_{h,\Delta t}$ *et* $(v_{\mathcal{M}_h,\Delta t} - v_{\mathcal{T}_h,\Delta t})_{h,\Delta t}$ *convergent fortement vers zéro dans* $L^2(Q_T)$ *quand* $h \to 0$.

Nous sommes en mesure d'annoncer le résultat principal de ce chapitre concernant la convergence du schéma non linéaire (1.20)–(1.21) vers une solution faible du problème continue (1.16)–(1.18) et quand les pas de discrétisations spatio-temporel tendent vers zéro.

Theorem 1.7. *Soit* $(\mathcal{T}_m)_{m\geq 1}$ *une suite de triangulations conformes de* Ω *telles que* $h_{\mathcal{T}_m} \to 0$ *quand* $m \to \infty$*, et soit* $(\Delta t_m)_{m\geq 1}$ *une suite de pas de temps telle que* $\Delta t_m \to 0$ *quand* $m \to \infty$*. Sous les hypothèses* (P1) *et* (P2)*, la solution discrète* $(u_{\mathcal{M}_m,\Delta t_m}, v_{\mathcal{M}_m,\Delta t_m})_m$ *converge fortement dans* $L^q(Q_{t_{\mathrm{f}}})$ *pour tout* $q \in [1, \infty)$ *vers une solution faible du problème continue* (1.16)–(1.18) *au sens de la définition* (1.1) *quand* $m \to \infty$.

La résolution du schéma (1.20)–(1.21) nécessite toujours la résolution d'un système non linéaire. Pour cela, la méthode de Newton est employée pour approcher la solution U^{n+1} de l'équation non linéaire définie par (1.20). Cet algorithme est couplé par une méthode de bigradient conjugué pour résoudre les systèmes linéaires découlant de la méthode de Newton aussi bien que la solution V^{n+1} de l'équation linéaire définie par (1.21).

La dernière partie de ce chapitre est constituée de trois simulations numériques effectuées avec le schéma numérique (1.20)–(1.21). Ces simulations ont pour but de vérifier l'efficacité du schéma proposé pour capturer la génération des patterns spatiaux pour le modèle de chimiotaxie sous l'effet du remplissage du volume. La première simulation consiste à prendre des tenseurs isotropes (c-à-d proportionnels à la matrice d'identité) ; ce cas d'étude est important afin de vérifier que les résultats effectués avec le schéma (1.20)–(1.21) sont cohérents avec ceux effectués en utilisant le schéma de volumes finis classique. La deuxième simulation consiste à prendre Λ comme étant un tenseur anisotrope diagonale dont les éléments diagonaux sont distincts, et à prendre une matrice isotrope pour D. Ce test montre que le modèle (1.16)–(1.18) génère des patterns spatiaux qui se propagent dans la direction d'un axe faisant avec l'axe des abscisses un angle identique à celui calculé en utilisant la matrice de passage pour le tenseur Λ. La dernière simulation consiste aussi à prendre une matrice isotrope pour D et à prendre une matrice anisotrope hétérogène pour Λ. Pour ce dernier test, nous obtenons également des patterns spatiaux pour lesquels la direction de propagation est cohérente avec l'angle calculé par la matrice de passage (matrice de rotation) du tenseur Λ.

1.2.3 Chapitre 4 : Un schéma CVFE non linéaire pour le modèle de Keller–Segel modifié

Dans ce chapitre, nous nous sommes intéressés à un modèle généralisé de Keller–Segel anisotrope et dégénéré. Le modèle ressemble à celui de la section précédente, pour lequel nous avons ajouté des tenseurs pour les termes de diffusion et pour le terme de convection. Plus précisément, nous considérons la formulation suivante :

$$\begin{cases} \partial_t u - \operatorname{div}\left(\Lambda\left(\mathbf{x}\right) a\left(u\right) \nabla u - \Lambda\left(\mathbf{x}\right) \chi\left(u\right) \nabla v\right) = f\left(u\right) & \text{dans } Q_{t_\mathrm{f}} = \Omega \times (0, t_\mathrm{f}), \\ \partial_t v - \operatorname{div}\left(D\left(\mathbf{x}\right) \nabla v\right) = g\left(u, v\right) & \text{dans } Q_{t_\mathrm{f}} = \Omega \times (0, t_\mathrm{f}), \end{cases} \tag{1.22}$$

avec les conditions aux limites données sur le bord $\Sigma_{t_\mathrm{f}} := \partial\Omega \times (0, t_\mathrm{f})$ par

$$\left(\Lambda\left(\mathbf{x}\right) a\left(u\right) \nabla u - \Lambda\left(\mathbf{x}\right) \chi\left(u\right) \nabla v\right) \cdot \mathbf{n} = 0, \qquad D\left(\mathbf{x}\right) \nabla v \cdot \mathbf{n} = 0, \tag{1.23}$$

et les conditions initiales données par :

$$u\left(\mathbf{x}, 0\right) = u_0\left(\mathbf{x}\right), \quad v\left(\mathbf{x}, 0\right) = v_0\left(\mathbf{x}\right), \quad \mathbf{x} \in \Omega. \tag{1.24}$$

Le domaine occupé par les cellules et les substances chimiques est toujours noté Ω, un ensemble ouvert et borné de $\mathbb{R}^2$. Nous donnons également les hypothèses portant sur la diffusion de la densité cellulaire dégénérée a, la sensibilité chimiotactique χ, les tenseurs de diffusion Λ et D, les cinétiques et les conditions initiales.

(A1) $a : [0, 1] \longrightarrow \mathbb{R}$ est une fonction continue telle que $a\left(0\right) = a\left(1\right) = 0$ et $a(u) > 0$ pour tout $0 < u < 1$.

(A2) $\chi : [0, 1] \longrightarrow \mathbb{R}$ est une fonction dérivable telle que $\chi(0) = \chi(1) = 0$ et $\chi(u) > 0$ pour tout $0 < u < 1$. En outre, il existe une fonction $\mu \in \mathcal{C}^0\left([0, 1], \mathbb{R}\right)$ telle que $\mu\left(u\right) = \frac{\chi(u)}{a(u)}$, $\mu\left(0\right) = \mu\left(1\right) = 0$ et $\mu\left(u\right) > 0$ pour tout $0 < u < 1$.

(A3) Les tenseurs de diffusion Λ et D sont deux tenseurs symétriques bornés, uniformément positifs sur Ω, c-à-d tels que :
$\forall \mathbf{w} \neq 0,$ il existe deux constantes strictement positives T_- et T_+ telles que $0 < T_- |\mathbf{w}|^2 \leq \langle T(\mathbf{x})\mathbf{w}, \mathbf{w}\rangle \leq T_+ |\mathbf{w}|^2 < \infty, T = \Lambda$ ou D.

(A4) La fonction $f : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction continue telle que $f(0) \geq 0$ et $f(1) \leq 0$.

(A5) Les conditions initiales u_0 et v_0 sont deux fonctions dans $L^\infty(\Omega)$ telles que $0 \leq u_0 \leq 1$ et $v_0 \geq 0$.

Notons $\xi : [0,1] \longrightarrow \mathbb{R}$ la fonction Lipschitzienne et croissante définie par $\xi(u) = \int_0^u \sqrt{a(s)}\,\mathrm{d}s$, pour tout $u \in \mathbb{R}$.

Definition 1.8 (Solution faible)**.** Sous les hypothèses (A1)–(A5), le couple (u, v) est dit solution faible du système (1.22)–(1.24) si il vérifie :

$$
\begin{aligned}
&0 \leq u(\mathbf{x},t) \leq 1,\ v(\mathbf{x},t) \geq 0 \text{ pour presque tout } (\mathbf{x},t) \in Q_{t_\mathrm{f}},\\
&\xi(u) \in L^2\left(0,t_\mathrm{f}; H^1(\Omega)\right),\\
&v \in L^\infty(Q_{t_\mathrm{f}}) \cap L^2\left(0,t_\mathrm{f}; H^1(\Omega)\right),
\end{aligned}
$$

et pour tout $\varphi, \psi \in \mathcal{D}\left(\overline{\Omega} \times [0,t_\mathrm{f})\right)$

$$
\begin{aligned}
-\int_\Omega u_0(\mathbf{x})\varphi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} - \iint_{Q_{t_\mathrm{f}}} u\partial_t\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \iint_{Q_{t_\mathrm{f}}} \sqrt{a(u)}\Lambda(\mathbf{x})\nabla\xi(u)\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
- \iint_{Q_{t_\mathrm{f}}} \Lambda(\mathbf{x})\chi(u)\nabla v\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \iint_{Q_{t_\mathrm{f}}} f(u)\varphi(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t,\\
-\int_\Omega v_0(\mathbf{x})\psi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} - \iint_{Q_{t_\mathrm{f}}} v\partial_t\psi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
+ \iint_{Q_{t_\mathrm{f}}} D(\mathbf{x})\nabla v\cdot\nabla\psi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \iint_{Q_{t_\mathrm{f}}} g(u,v)\psi(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.
\end{aligned}
$$

Dans ce chapitre, nous construisons et analysons un schéma numérique pour la discrétisation du système (1.22)–(1.24). Nous adoptons le principe du schéma non linéaire CVFE proposé dans la section précédente pour la construction du schéma. Dans ce dernier schéma, les degrés de liberté sont affectés aux sommets d'un maillage triangulaire primal, comme dans la méthode des éléments finis, tandis que les équations de bilan sont discrétisées sur un maillage dual spécifique (le maillage barycentrique dual appelé encore le maillage de Donald) en utilisant les flux de diffusion fournis par la reconstruction des éléments finis sur le maillage triangulaire primal. Plus particulièrement, les termes de diffusion sont approchés par l'intermédiaire des flux de diffusion découlant de la reconstruction des éléments finis et en utilisant un schéma de Godunov. Par contre, le terme de convection est approché par le moyen d'un schéma amont original ; en effet, la fonction χ est définie comme étant le produit de deux fonctions μ et a, le schéma proposé consiste à prendre le schéma amont classique d'une part, pour la discrétisation de la fonction μ qui sera approchée par l'intermédiaire d'une fonction flux numérique et à prendre le schéma de Godunov d'autre part, pour approcher la fonction a qui intervient dans la fonction de la sensibilité chimiotactique χ. Les approximations utilisées dans ce schéma sont cruciales afin d'assurer le principe de maximum, la stabilité et la convergence du schéma sans aucune restriction sur les coefficients de transmissibilité.

Comme dans la section précédente, la discrétisation du système (1.22)–(1.24) exige la construction de deux types d'approximations : l'approximation éléments finis sur un maillage triangulaire primal et une approximation volumes finis sur le maillage barycentrique correspondant.

Nous construisons le maillage primal $\mathcal{T}$ à partir d'un nombre fini de triangles disjoints qui forment une partition du domaine Ω supposé polygonal ; nous avons alors $\bigcup_{T\in\mathcal{T}}\overline{T} = \overline{\Omega}$, et $T \cap T' = \emptyset$ si $T \neq T'$. Nous notons par $\mathcal{V}$ l'ensemble de tous les sommets de la triangulation conforme $\mathcal{T}$ et par $\mathcal{E}$ l'ensemble de toutes les arêtes de la triangulation $\mathcal{T}$. Pour chaque sommet $K \in \mathcal{V}$ (localisé à la position $\mathbf{x}_K$), nous notons par $\mathcal{E}_K$ le sous-ensemble de $\mathcal{E}$ constitué des arêtes admettant le sommet K comme une extrémité. Une arête joignant deux sommets K et L est notée par σ_{KL}.

Pour la construction du maillage barycentrique dual, nous notons par T_K l'ensemble de tous les triangles admettant K comme un sommet. Il existe un volume de contrôle unique ω_K associé au sommet K, il est construit autour de ce sommet en joignant les centres de gravité $\mathbf{x}_T$ des triangles $T \in \mathcal{T}_K$ avec les milieux $\mathbf{x}_\sigma$ des arêtes $\sigma \in \mathcal{E}_K$ (voir Figure 1.6). Nous déterminons les constantes

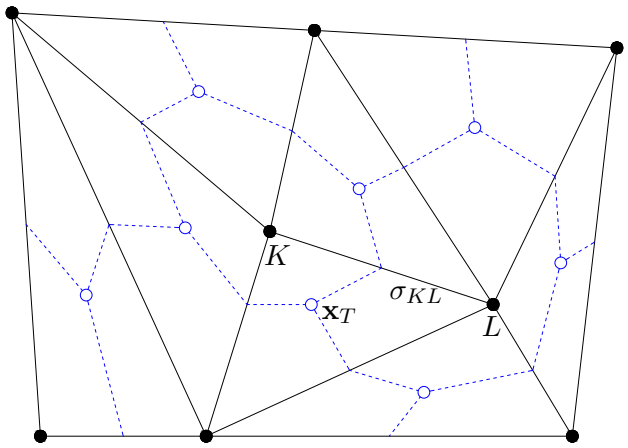


FIGURE 1.6 – Le maillage triangulaire primal $\mathcal{T}$ et le maillage de Donald dual $\mathcal{M}$.

$h_\mathcal{T}$ et $\theta_\mathcal{T}$ comme étant la taille et la régularité du maillage Ω respectivement. Elles sont définies par :

$$h_\mathcal{T} = \max_{T\in\mathcal{T}}, \qquad \theta_\mathcal{T} = \max_{T\in\mathcal{T}}\frac{h_T}{\rho_T},$$

où h_T est le diamètre du triangle T et ρ_T est le diamètre du cercle circonscrit au triangle T.

Nous notons $\mathcal{H}_\mathcal{T}$ l'espace usuel des éléments finis c-à-d l'espace $\mathbb{P}_1$-discret formé des fonctions continues et affines par morceaux sur le maillage triangulaire primal

$$\mathcal{H}_\mathcal{T} = \{\phi \in \mathcal{C}^0\left(\overline{\Omega}\right) ; \phi_{|_T} \in \mathbb{P}_1\left(\mathbb{R}\right), \forall T \in \mathcal{T}\} \subset H^1\left(\Omega\right),$$

et par $(\varphi_K)_{K\in\mathcal{V}}$ sa base canonique définie par $\varphi_K\left(\mathbf{x}_L\right) = \delta_{KL}$ (δ est le symbole de Kronecker). En outre, nous considérons l'espace discret de volumes finis $\mathcal{X}_\mathcal{M}$ formé des fonctions constantes par morceaux sur le maillage barycentrique dual

$$\mathcal{X}_\mathcal{M} = \{\phi : \Omega \longrightarrow \overline{\mathbb{R}},\ \phi_{|_{\omega_K}} \text{ est une constante}, \forall K \in \mathcal{V}\}.$$

Dans le but de simplifier les notations dans le chapitre, nous limitons notre étude au cas d'une discrétisation uniforme en temps. Pour cela, nous notons $\Delta t = t_\mathrm{f}/(N+1)$ le pas de temps uniforme et $t_n = n\Delta t$ de sorte que $t^0 = 0$ et $t_{N+1} = t_\mathrm{f}$. Une fois l'intervalle de temps discrétisé, nous introduisons les espaces discrets spatiaux et temporels définis par :

$$\mathcal{H}_{\mathcal{T},\Delta t} = \{\phi \in L^2\left(0, t_\mathrm{f}; H^1\left(\Omega\right)\right) ; \phi\left(\cdot, t\right) = \phi\left(\cdot, t_{n+1}\right) \in \mathcal{H}_\mathcal{T},\ \forall t \in (t_n, t_{n+1}], 0 \leq n \leq N\},$$
$$\mathcal{X}_{\mathcal{M},\Delta t} = \{\phi \in L^\infty\left(Q_{t_\mathrm{f}}\right) ; \phi\left(\cdot, t\right) = \phi\left(\cdot, t_{n+1}\right) \in \mathcal{X}_\mathcal{M},\ \forall t \in (t_n, t_{n+1}], 0 \leq n \leq N\}.$$

Pour un vecteur donné $(u_K^n)_{n\in\{0,\cdots,N+1\},K\in\mathcal{V}}$ (resp. $(v_K^n)_{n\in\{0,\cdots,N+1\},K\in\mathcal{V}}$), il existe une unique fonction $u_{\mathcal{T},\Delta t}\in\mathcal{H}_{\mathcal{T},\Delta t}$ (resp. $v_{\mathcal{T},\Delta t}\in\mathcal{H}_{\mathcal{T},\Delta t}$) et une unique fonction $u_{\mathcal{M},\Delta t}\in\mathcal{X}_{\mathcal{M},\Delta t}$ (resp. $v_{\mathcal{M},\Delta t}\in\mathcal{X}_{\mathcal{M},\Delta t}$) telles que :

$$\begin{aligned} u_{\mathcal{T},\Delta t}(\mathbf{x}_K,t_{n+1}) = u_{\mathcal{M},\Delta t}(\mathbf{x}_K,t_{n+1}) = u_K^{n+1}, \qquad \forall K\in\mathcal{V},\ \forall n\in\{0,\dots,N\},\\ v_{\mathcal{T},\Delta t}(\mathbf{x}_K,t_{n+1}) = v_{\mathcal{M},\Delta t}(\mathbf{x}_K,t_{n+1}) = v_K^{n+1}, \qquad \forall K\in\mathcal{V},\ \forall n\in\{0,\dots,N\}. \end{aligned}$$

Soit m_K la mesure de Lebesgue dans $\mathbb{R}^2$ du volume de contrôle ω_K. Pour chaque couple $(K,L)\in\mathcal{V}^2$, nous notons Λ_{KL} et D_{KL} les coefficients de transmissibilité définis par :

$$\mathbf{T}_{KL} = \int_\Omega \mathbf{T}(\mathbf{x})\,\nabla\varphi_K(\mathbf{x})\cdot\nabla\varphi_L(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad \mathbf{T}\equiv\Lambda \text{ ou } D. \tag{1.25}$$

Le schéma proposé consiste à intégrer le système (1.22) sur $\omega_K\times[t_n,t_{n+1}]$ avec $K\in\mathcal{V}$ et à utiliser le théorème de Green–Gauss pour passer à l'interface de chaque volume dual.

Discrétisation de la deuxième équation du système (1.22) Nous détaillons la discrétisation de la première équation du système (1.22). Le théorème de Green–Gauss donne :

$$\begin{aligned} \int_{t_n}^{t_{n+1}}\int_{\omega_K}\partial_t u\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \sum_{\sigma\subset\partial\omega_K}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda(\mathbf{x})\,a(u)\,\nabla u\cdot\eta_\sigma\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \\ + \sum_{\sigma\subset\partial\omega_K}\int_{t_n}^{t_{n+1}}\int_\sigma \mu(u)\,\Lambda(\mathbf{x})\,a(u)\,\nabla v\cdot\eta_\sigma\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_{\omega_K} f(u)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \end{aligned}$$

Le terme d'évolution en temps

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_{\omega_K}\partial_t u\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \approx \frac{1}{\Delta t}\int_{\omega_K}\left(u(\mathbf{x},t_{n+1})-u(\mathbf{x},t_n)\right)\mathrm{d}\mathbf{x} \approx m_K\frac{u_K^{n+1}-u_K^n}{\Delta t}.$$

Le terme de diffusion Nous voulons approcher le flux $\mathcal{F}(u) = -\Lambda(\mathbf{x})\,a(u)\,\nabla u\cdot\eta_\sigma$ sur l'interface σ (ici σ est l'interface partagée par ω_K avec son volume de contrôle dual voisin ω_L et appartenant au triangle T tel que $T\cap\omega_K\neq\emptyset\neq T\cap\omega_L$). Le schéma de Godunov consiste à approcher le flux $\mathcal{F}$ en utilisant une fonction flux numérique $\mathcal{F}_{KL}$ définie par :

$$\mathcal{F}_{KL}(u_K,u_L) = \begin{cases} \min\limits_{u\in[u_K,u_L]}\Lambda_{KL}(u_K-u_L)\,a(u), & \text{si } u_K\le u_L,\\ \max\limits_{u\in[u_L,u_K]}\Lambda_{KL}(u_K-u_L)\,a(u), & \text{si } u_L\le u_K, \end{cases}$$

où $\Lambda_{KL}(u_K-u_L)$ représente une approximation du flux de diffusion $-\Lambda\nabla u\cdot\eta_\sigma$ sur l'interface σ et provenant de la construction éléments finis (comme pour la section précédente) et où $\Lambda_{KL} = \sum_{K\in\mathcal{T}}\Lambda_{KL}^T$ avec $\Lambda_{KL}^T = \int_T\Lambda\nabla\varphi_K\cdot\nabla\varphi_L\mathrm{d}\mathbf{x}$.
En utilisant les propriétés classiques sur les fonctions Lipschitziennes min et max, nous obtenons

$$\mathcal{F}_{KL}(u_K,u_L) = \begin{cases} \Lambda_{KL}(u_K-u_L)\max\limits_{u\in[u_K,u_L]}a(u), & \text{si } \Lambda_{KL}\ge 0,\\ \Lambda_{KL}(u_K-u_L)\min\limits_{u\in[u_K,u_L]}a(u), & \text{si } \Lambda_{KL}\le 0,\\ \Lambda_{KL}(u_K-u_L)\max\limits_{u\in[u_L,u_K]}a(u), & \text{si } \Lambda_{KL}\ge 0,\\ \Lambda_{KL}(u_K-u_L)\min\limits_{u\in[u_L,u_K]}a(u), & \text{si } \Lambda_{KL}\le 0. \end{cases}$$

D'une manière simplifiée, nous écrivons $\mathcal{F}_{KL}\left(u_K, u_L\right) = \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right)$ à l'instant t_{n+1} avec

$$a_{KL}^{n+1} = \begin{cases} \max_{s \in I_{KL}^{n+1}} a(s) & \text{si } \Lambda_{KL} \geq 0, \\ \min_{s \in I_{KL}^{n+1}} a(s) & \text{si } \Lambda_{KL} < 0, \end{cases} \quad \text{et} \quad I_{KL}^{n+1} \begin{cases} [u_K, u_L], & \text{si } u_K \leq u_L, \\ [u_L, u_K], & \text{si } u_L \leq u_K. \end{cases}$$

Par conséquent, nous obtenons l'approximation suivante du terme de diffusion :

$$\begin{aligned} -\frac{1}{\Delta t} \sum_{\sigma \subset \partial\omega_K} \int_{t_n}^{t_{n+1}} \int_{\sigma} \Lambda\left(\mathbf{x}\right) a\left(u\right) \nabla u \cdot \eta_\sigma \, \mathrm{d}\sigma(\mathbf{x}) \, \mathrm{d}t &\approx \sum_{T \in \mathcal{T}} \sum_{\sigma \in \partial\omega_K \cap T} \Lambda_{KL}^T a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right) \\ &= \sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right). \end{aligned}$$

Le terme de convection La discrétisation du terme de convection ressemble à celle du terme de diffusion ; en effet, la fonction χ s'écrit comme le produit de deux fonctions μ et a. La fonction a est approchée en utilisant un schéma de Godunov, et ensuite nous utilisons un schéma décentré amont pour approcher le flux $\mu\left(u\right)\Lambda\left(\mathbf{x}\right)\nabla v \cdot \eta_\sigma$ à travers l'interface σ, ce qui nécessite l'introduction d'une fonction flux numérique vérifiant les propriétés (i)–(v) de la section 1.2.1. Nous donnons deux exemples sur l'approximation μ_{KL}^{n+1} de la fonction μ de telle sorte que la fonction flux numérique vérifie (i)–(v). Le premier exemple consiste à prendre le schéma d'Engquist-Osher et le deuxième consiste à prendre le schéma de Godunov. En particulier, nous avons :

- $\mu_{KL}^{n+1} = \begin{cases} \mu_\downarrow\left(u_K^{n+1}\right) + \mu_\uparrow\left(u_L^{n+1}\right), & \text{si } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \mu_\uparrow\left(u_K^{n+1}\right) + \mu_\downarrow\left(u_L^{n+1}\right), & \text{si } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0. \end{cases}$
  Les fonctions $\mu_\uparrow$ et $\mu_\downarrow$ sont données par :

$$\mu_\uparrow\left(z\right) := \int_0^z \left(\mu'\left(s\right)\right)^+ \mathrm{d}s, \qquad \mu_\downarrow\left(z\right) := -\int_0^z \left(\mu'\left(s\right)\right)^- \mathrm{d}s.$$

- $\mu_{KL}^{n+1} = \begin{cases} \max\limits_{[u_K^{n+1}, u_L^{n+1}]} \mu\left(u\right), & \text{si } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \min\limits_{[u_L^{n+1}, u_K^{n+1}]} \mu\left(u\right), & \text{si } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \max\limits_{[u_L^{n+1}, u_K^{n+1}]} \mu\left(u\right), & \text{si } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0, \\ \min\limits_{[u_K^{n+1}, u_L^{n+1}]} \mu\left(u\right), & \text{si } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0. \end{cases}$

Par conséquent, nous obtenons l'approximation suivante pour de terme de convection :

$$\frac{1}{\Delta t} \sum_{\sigma \subset \partial\omega_K} \int_{t_n}^{t_{n+1}} \int_{\sigma} \mu\left(u\right) \Lambda\left(\mathbf{x}\right) a(u) \nabla v \cdot \eta_\sigma \, \mathrm{d}\sigma(\mathbf{x}) \, \mathrm{d}t \approx -\sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL} \mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right).$$

Discrétisation de la deuxième équation du système (1.22) Dans le but de discrétiser la deuxième équation du système (1.22), nous sommes amenés à introduire les fonctions suivantes $\eta\left(v\right)$, $p\left(v\right)$,

$\Gamma(v)$ et $\phi(v)$ définies par :

$$\eta(v) = \max(0, \min(v, 1)), \quad p(v) = \int_1^v \frac{1}{\eta(s)}\,\mathrm{d}s = \begin{cases} \ln(v) & \text{if } v \in (0,1), \\ v-1 & \text{if } v \geq 1, \end{cases}$$
$$\Gamma(v) = \int_1^v p(s)\,\mathrm{d}s = \begin{cases} v\ln(v) - v + 1 & \text{if } v \in [0,1), \\ \frac{(v-1)^2}{2} & \text{if } v \geq 1, \end{cases}$$
$$\phi(v) = \int_0^v \frac{1}{\sqrt{\eta(s)}}\,\mathrm{d}s = \begin{cases} \frac{\sqrt{v}-1}{2} & \text{if } v \in [0,1), \\ v-1 & \text{if } v \geq 1. \end{cases}$$

En utilisant ces fonctions, nous obtenons une approximation pour le terme de diffusion similaire à celle obtenue pour la première équation.

En résumé, le schéma CVFE non linéaire proposé pour la discrétisation du problème (1.22)–(1.24) consiste à chercher $U = \left(u_K^{n+1}\right)_{K\in\mathcal{V},n\in[0,\ldots,N+1]}$ et $V = \left(v_K^{n+1}\right)_{K\in\mathcal{V},n\in[0,\ldots,N+1]}$ solution de

$$u_{\mathcal{M}}^0(\mathbf{x}) = u_K^0 = \frac{1}{m_K}\int_{\omega_K} u_0(\mathbf{y})\,\mathrm{d}\mathbf{y}, \qquad v_{\mathcal{M}}^0(\mathbf{x}) = v_K^0 = \frac{1}{m_K}\int_{\omega_K} v_0(\mathbf{y})\,\mathrm{d}\mathbf{y}. \tag{1.26}$$

et pour tout $n \in \{0,\ldots,N\}$

$$\begin{aligned} m_K\frac{u_K^{n+1}-u_K^n}{\Delta t} &+ \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right) \\ &- \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL}\mu_{KL}^{n+1}a_{KL}^{n+1}\left(v_K^{n+1}-v_L^{n+1}\right) = m_K f\left(u_K^{n+1}\right), \\ m_K\frac{v_K^{n+1}-v_K^n}{\Delta t} &+ \sum_{\sigma_{KL}\in\mathcal{E}_K} D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right)-p\left(v_L^{n+1}\right)\right) = m_K\left(\alpha u_K^n - \beta v_K^{n+1}\right). \end{aligned} \tag{1.27}$$

Conservativité du schéma Par construction, le schéma (1.26)–(1.27) est un schéma conservatif ; en effet, en notant $F_{KL}^{n+1} = \Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right) - \Lambda_{KL}\mu_{KL}^{n+1}a_{KL}^{n+1}\left(v_K^{n+1}-v_L^{n+1}\right)$ et $\Phi_{KL}^{n+1} = D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right)-p\left(v_L^{n+1}\right)\right)$, nous avons $a_{KL}^{n+1} = a_{LK}^{n+1}$, $\eta_{KL}^{n+1} = \eta_{LK}^{n+1}$ et $\mu_{KL}^{n+1} = \mu_{LK}^{n+1}$. Par conséquent, nous obtenons la forme conservative locale

$$\begin{cases} F_{KL}^{n+1} + F_{LK}^{n+1} = 0 = \Phi_{KL}^{n+1} + \Phi_{LK}^{n+1}, & \text{pour tout } \sigma_{KL} \in \mathcal{E}, \\ m_K\dfrac{u_K^{n+1}-u_K^n}{\Delta t} + \displaystyle\sum_{\sigma_{KL}\in\mathcal{E}_K} F_{KL}^{n+1} = f\left(u_K^{n+1}\right)m_K, & \text{pour tout } K \in \mathcal{V}, \\ m_K\dfrac{v_K^{n+1}-v_K^n}{\Delta t} + \displaystyle\sum_{\sigma_{KL}\in\mathcal{E}_K} \Phi_{KL}^{n+1} = g\left(u_K^n, v_K^{n+1}\right)m_K, & \text{pour tout } K \in \mathcal{V}. \end{cases}$$

Afin d'établir l'existence d'une solution et la convergence du schéma non linéaire (1.26)–(1.27), nous utilisons des propriétés discrètes ainsi que des estimations *a priori*. Toutes ces propriétés seront détaillées dans le chapitre. Nous commençons par la propriété suivante

$$\int_\Omega \mathbf{T}(\mathbf{x})\nabla u_{\mathcal{T}}\cdot\nabla v_{\mathcal{T}}\mathrm{d}\mathbf{x} = \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(u_K-u_L)(v_K-v_L), \quad \mathbf{T}(\mathbf{x}) = \Lambda(\mathbf{x}) \text{ ou } D(\mathbf{x}).$$

En effet, notons que nous avons $\mathbf{T}_{KK} = -\sum_{L\neq K} \mathbf{T}_{KL}$. Par conséquent,

$$\begin{aligned}\int_\Omega \mathbf{T}(\mathbf{x})\nabla u_\mathcal{T}\cdot\nabla v_\mathcal{T}\mathrm{d}\mathbf{x} &= \sum_{T\in\mathcal{T}}\int_T \mathbf{T}(\mathbf{x})\nabla u_\mathcal{T}\cdot\nabla v_\mathcal{T}\mathrm{d}\mathbf{x}\\ &= -\sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}u_K v_L - \sum_{K\in\mathcal{V}}\mathbf{T}_{KK}u_K v_K = \sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}(v_K - v_L)u_K\\ &= \sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}(v_K - v_L)u_K = \sum_{\sigma_{KL}\in\mathcal{E}}\mathbf{T}_{KL}(v_K - v_L)(u_K - u_L).\end{aligned}$$

Nous donnons les deux lemmes suivants dont la démonstration est une conséquence directe de la propriété précédente ainsi que de la définition des fonctions ξ et ϕ.

Proposition 1.9. *Notons $\xi_{\mathcal{T},\Delta t}$ (resp. $\phi_{\mathcal{T},\Delta t}$) l'unique fonction de $\mathcal{H}_{\mathcal{T},\Delta t}$ avec des valeurs nodales $\left(\xi\left(u_K^{n+1}\right)\right)\in\mathbb{R}^{(N+1)\#\mathcal{V}}$ (resp. $\left(\phi\left(v_K^{n+1}\right)\right)\in\mathbb{R}^{(N+1)\#\mathcal{V}}$). Nous avons alors*

$$\begin{aligned}&\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2\\ &\qquad\geq\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\Lambda_{KL}\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2=\iint_{Q_{t_\mathrm{f}}}\Lambda\nabla\xi_{\mathcal{T},\Delta t}\cdot\nabla\xi_{\mathcal{T},\Delta t}\mathrm{d}\mathbf{x}\mathrm{d}t,\end{aligned}$$

et

$$\begin{aligned}&\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1})-p(v_L^{n+1})\right)^2\\ &\qquad\geq\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}\left(\phi\left(v_K^{n+1}\right)-\phi\left(v_L^{n+1}\right)\right)^2=\iint_{Q_{t_\mathrm{f}}}D\nabla\phi_{\mathcal{T},\Delta t}\cdot\nabla\phi_{\mathcal{T},\Delta t}\mathrm{d}\mathbf{x}\mathrm{d}t.\end{aligned}$$

Proposition 1.10. *Il existe une constante $C_1>0$ qui dépend seulement de Λ et $\theta_\mathcal{T}$ telle que*

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|\Lambda_{KL}|\,a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2\leq C_1\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2.$$

$$\begin{aligned}&\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|D_{KL}|\,\eta_{KL}^{n+1}\left(p(v_K^{n+1})-p(v_L^{n+1})\right)^2\\ &\qquad\leq C_1\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1})-p(v_L^{n+1})\right)^2.\end{aligned}$$

Estimations discrètes, existence et convergence du schéma

Nous donnons les estimations *a priori* sur les solutions discrètes du schéma non linéaire (1.26)–(1.27). Nous présentons tout d'abord le principe de maximum discret :

Proposition 1.11. *Soit $\left(u_K^{n+1},v_K^{n+1}\right)_{K\in\mathcal{V},\,n\in\{0,\ldots,N\}}$ une solution du schéma non linéaire (1.26)–(1.27). Alors, pour tout $K\in\mathcal{V}_h$, et pour tout $n\in\{0,\ldots,N+1\}$, nous avons $0\leq u_K^n\leq 1$ et $v_K^n\geq 0$.*

La preuve de cette proposition est classique. Nous prenons un volume de contrôle dual ω_K tel que $u_K^{n+1} = \min\{u_L^{n+1}\}_{L\in\mathcal{V}}$, nous multiplions la première équation du système (1.27) par $-\left(u_K^{n+1}\right)^-$ et nous faisons la sommation sur $K \in \mathcal{V}$. Nous obtenons la positivité de u_K^{n+1} en utilisant le prolongement par continuité des fonctions μ et f pour $u \leq 0$ et en remarquant que $a_{KL}^{n+1} = 0$ quand $\Lambda_{KL} \geq 0$, et donc $a_{KL}^{n+1}\left(\Lambda_{KL}\right)^+\left(u_K^{n+1} - u_L^{n+1}\right)\left(u_K^{n+1}\right)^- \geq 0$ et par suite $\mu_{KL}^{n+1}\Lambda_{KL}^+\left(v_K^{n+1} - v_L^{n+1}\right)^-\left(u_K^{n+1}\right)^- = 0$. D'une manière similaire, nous multiplions l'équation par $\left(u_K^{n+1} - 1\right)^+$ pour montrer que $u_K^{n+1} \leq 1$. La positivité de v_K^n découle directement de la positivité de u_K^n pour tout $K \in \mathcal{V}$ et tout $n \in \{0, \ldots, N+1\}$.

Estimations *a priori* Nous avons les estimations d'énergie suivantes :

Proposition 1.12. *Pour tout $n \geq 0$, il existe une constante $C > 0$ indépendante de h et telle que*

$$\sum_{K\in\mathcal{V}} m_K \Gamma\left(v_K^{n+1}\right) + \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} D_{KL}\left(\phi(v_K^{n+1}) - \phi(v_L^{n+1})\right)^2 \leq \sum_{K\in\mathcal{V}} m_K \Gamma(v_K^{n+1}) + \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2 \leq C.$$

Proposition 1.13. *Il existe une constante $C > 0$ indépendante de h et telle que*

$$\iint_{Q_{t_\mathrm{f}}} \Lambda\left(\mathbf{x}\right)\nabla v_{\mathcal{T},\Delta t}\left(\mathbf{x},t\right)\cdot\nabla v_{\mathcal{T},\Delta t}\left(\mathbf{x},t\right)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right)^2 \leq C.$$

En utilisant les propositions précédentes, nous obtenons des estimations analogues à celles données par la proposition 1.12. Spécifiquement, nous avons

Proposition 1.14. *Pour tout $n \geq 0$, il existe une constante $C > 0$ indépendante de h et telle que*

$$\sum_{K\in\mathcal{V}} m_K\left(u_K^{n+1}\right)^2 + \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL}\left(\xi(u_K^{n+1}) - \xi(u_L^{n+1})\right)^2 \leq \sum_{K\in\mathcal{V}} m_K\left(u_K^{n+1}\right)^2 + \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right)^2 \leq C.$$

Dans tout ce qui suite, nous notons par $u_{\mathcal{M},\Delta t}$ et $v_{\mathcal{M},\Delta t}$ les éléments uniques de l'espace $\mathcal{X}_{\mathcal{M},\Delta t}$ tels que

$$u_{\mathcal{M},\Delta t}\left(\mathbf{x}_K,t\right) = u_K^{n+1}, \quad v_{\mathcal{M},\Delta t}\left(\mathbf{x}_K,t\right) = v_K^{n+1}, \quad \forall K \in \mathcal{V}, \forall n \geq 0.$$

Nous montrons une estimation portant sur un raffinement amélioré de $v_{\mathcal{M},\Delta t}$. Cette estimation prétend que la fonction $v_{\mathcal{M},\Delta t}$ est ou bien une constante égale à zéro ou bien $v_{\mathcal{M},\Delta t} > 0$.

Proposition 1.15. *Supposons que $\int_\Omega u_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$ ou $\int_\Omega v_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$. Alors il existe une constante r_h qui dépend des données et aussi de Δt et du maillage $\mathcal{T}$ telle que :*

$$v_K^{n+1} \geq r_h, \quad \forall K \in \mathcal{V}, \forall n \in \{0, \ldots, N\}.$$

En utilisant les propositions 1.12–1.15, nous nous appuyons sur l'argument de degré topologique pour obtenir le résultat d'existence d'une solution discrète.

Proposition 1.16 (Existence d'une solution discrète)**.** *Soit $\left(u_K^n, v_K^n\right)_{K\in\mathcal{V}}$ un vecteur donné tel que $u_{\mathcal{M},\Delta t}(\cdot,t_n)$ et $v_{\mathcal{M},\Delta t}(\cdot,t_n)$ sont positives. Alors, le schéma non linéaire (1.26)–(1.26) admet au moins une solution $\left(u_K^{n+1}, v_K^{n+1}\right)_{K\in\mathcal{V}}$. En outre, $u_{\mathcal{M},\Delta t}(\cdot,t_{n+1})$ and $v_{\mathcal{M},\Delta t}(\cdot,t_{n+1})$ sont positives.*

Le résultat principal de ce travail est le théorème suivant :

Theorem 1.17. *Soient $(\mathcal{T}_m)_{m\geq 1}$ une suite de triangulations conformes de Ω telle que $h_{\mathcal{T}_m} \to 0$ quand $m \to \infty$, et $(\Delta t_m)_{m\geq 1}$ une suite de pas de temps telle que $\Delta t_m \to 0$ quand $m \to \infty$. Pour tout $q \in [1, \infty)$, la solution discrète $(u_{\mathcal{M}_m,\Delta t_m}, v_{\mathcal{M}_m,\Delta t_m})_m$ converge faiblement dans $L^q(Q_{t_f})$ vers la solution faible du problème continue (1.22)–(1.24) quand $m \to \infty$.*

Afin d'obtenir le résultat de convergence, nous utilisons le critère de compacité de Kolmogorov en nous appuyant sur certaines estimations sur les translations en temps et en espace des solutions discrètes. Ensuite, nous identifions la solution limite avec la solution faible du problème continue (1.22)–(1.24) au sens de la définition 1.8.

La dernière partie de ce chapitre est dédiée à faire des simulations numériques en dimension 2 afin de vérifier l'efficacité du schéma proposé pour approcher les problèmes paraboliques non linéaires anisotropes modélisant la chimiotaxie.

1.2.4 Chapitre 5 : Analyse d'une équation parabolique dégénérée modélisant la chimiotaxie ou les fluides compressibles en milieu poreux

Dans ce chapitre, nous nous sommes intéressés à une équation parabolique dégénérée et non linéaire découlant, soit de la modélisation d'un fluide compressible, soit de la modélisation de la chimiotaxie. L'équation faisons l'objet d'étude dans ce chapitre est donnée, dans $Q_T = \Omega \times (0, T)$, par l'équation parabolique suivante :

$$\partial_t u - \operatorname{div}\left(a(u)\nabla u - f(u)\mathbf{V}\right) - g(u)\operatorname{div}(\mathbf{V}) + \gamma a(u)\nabla u \cdot \tilde{\mathbf{V}} = 0. \tag{1.28}$$

Cette équation découle, soit de la modélisation de la chimiotaxie-fluide (avec $\tilde{\mathbf{V}}$ est la vitesse du fluide transportant la densité cellulaire u), soit de la loi de conservation de la masse du fluide noté u (et dans ce cas $\mathbf{V} = \tilde{\mathbf{V}}$). Le domaine occupé par le fluide est noté Ω, qui est un ensemble ouvert et borné de $\mathbb{R}^d$, $d = 2, 3$. T représente un instant fixé dans le temps, pour lequel nous posons $\Sigma_T = \partial\Omega \times (0, T)$. À l'équation (1.28), nous ajoutons la condition initiale donnée par :

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \text{dans } \Omega, \tag{1.29}$$

pour tout $\mathbf{x} \in \Omega$, et les conditions aux limites sur le bord Σ_T données par :

$$u(\mathbf{x}, t) = 0, \quad \text{sur } \Sigma_T, \tag{1.30}$$

Nous introduisons les hypothèses classiques en milieu poreux portant sur le système : (1.28)–(1.30)

(H1) Les champs de vitesse $\mathbf{V}$ et $\tilde{\mathbf{V}}$ sont deux fonctions mesurables dans $(L^\infty(\Omega))^d$. En outre, le champ de vitesse $\mathbf{V}$ vérifie la condition suivante :

$$\mathbf{V} \cdot \mathbf{n} \leq 0, \quad \text{sur } \Sigma_T,$$

où $\mathbf{n}$ est le vecteur normal unitaire à $\partial\Omega$ dirigé vers l'extérieur de Ω.

(H2) f est une fonction dérivable dans $[0, 1]$ et $g \in \mathcal{C}^1([0, 1])$ vérifiant

$$g(0) = f(0) = 0, \quad f(1) = g(1) = 1, \text{ et } g'(u) > 0 \quad \forall u \in [0, 1].$$

(H3) La condition initiale u_0 satisfait $u_0 \in L^2(\Omega)$ et $0 \leq u_0(\mathbf{x}) \leq 1$ pour presque tout $\mathbf{x} \in \Omega$.

Une difficulté majeure concernant le système (1.28)–(1.30) est la possibilité de dégénérescence du terme de diffusion. Nous donnons trois hypothèses de dégénérescence de la fonction de dissipation a qui correspondent à des situations physiques distinctes :

(H4a) $a \in \mathcal{C}^1([0,1],\mathbb{R})$, $a(u) > 0$ pour tout $0 < u < 1$, $a(0) > 0$, $a(1) = 0$.
De plus, il existe $a_0 > 0$, $0 < r_2 \leq 2$, $u_* < 1$, m_1 et $M_1 > 0$ tels que
$a(u) \geq a_0$ pour tout $0 \leq u \leq u_*$,
$m_1 (1-u)^{r_2} \leq a(u) \leq M_1 (1-u)^{r_2}$, pour tout $u_* \leq u \leq 1$. En outre, il existe c_1, $c_2 > 0$ tels que : $c_1 (1-u)^{-1} \leq (f(u) - g(u))^{-1} \leq c_2 (1-u)^{-1}$, pour tout $u_* \leq u < 1$.

(H4b) $a \in \mathcal{C}^1([0,1],\mathbb{R})$, $a(u) > 0$ pour tout $0 < u < 1$, $a(0) = 0$, $a(1) > 0$,
De plus, il existe $r_1 > 0$, m_1 et $M_1 > 0$ tels que
$m_1 r_1 u^{r_1 - 1} \leq a'(u) \leq M_1 r_1 u^{r_1 - 1}$, pour tout $0 \leq u \leq 1$. En outre, il existe une constant $C > 0$ telle que : $|f(u) - g(u)| \leq Cu$ pour tout $0 \leq u \leq 1$.

(H4c) $a \in \mathcal{C}^1([0,1],\mathbb{R})$, $a(u) > 0$ pour tout $0 < u < 1$, $a(0) = 0$, $a(1) = 0$,
De plus, il existe $r_1 > 0$, $r_2 > 0$, $u_* < 1$, m_1 et $M_1 > 0$ tels que
$m_1 r_1 u^{r_1 - 1} \leq a'(u) \leq M_1 r_1 u^{r_1 - 1}$, pour tout $0 \leq u \leq u_*$,
$-r_2 M_1 (1-u)^{r_2 - 1} \leq a'(u) \leq -r_2 m_1 (1-u)^{r_2 - 1}$, pour tout $u_* \leq u \leq 1$. En outre, il existe c_1, c_2, $C > 0$ tels que : $|f(u) - g(u)| \leq Cu$ pour tout $0 \leq u \leq u_*$ et $c_1 (1-u)^{-1} \leq (f(u) - g(u))^{-1} \leq c_2 (1-u)^{-1}$, pour tout $u_* \leq u < 1$.

Dans tout ce que suit, nous supposons que $\gamma = 1$ et que $\tilde{\mathbf{V}} = \mathbf{V}$. Nous établissons, selon les suppositions (H4a), (H4b) ou (H4c) trois résultats d'existence. Le premier résultat concerne l'existence des solutions du système (1.28)–(1.30) dans un sens classique, alors que les deux autres concernent l'existence des solutions dans un sens affaibli par rapport à la formulation classique.

Le premier résultat : les solutions faibles et classiques Nous définissons la fonction $k \in \mathcal{C}^1([0,1))$ par :

$$\begin{aligned} &k(u) = u, && \text{if } 0 \leq u \leq u_*, \\ &k'(u) = (f(u) - g(u))^{-1} g'(u)\, k(u), && \text{if } u_* \leq u < 1. \end{aligned} \tag{1.31}$$

Nous notons L la primitive de la fonction k définie par $L(u) = \int_0^u k(\tau)\, \mathrm{d}\tau$, pour tout $0 \leq u < 1$.

Definition 1.18. Sous les hypothèses (H1)–(H3) et (H4a), et pour une donnée initiale u_0 satisfaisant $L(u_0) \in L^1(\Omega)$, nous disons que u est une solution faible classique du système (1.28)–(1.30) si u vérifie

$$\begin{aligned} &0 \leq u(\mathbf{x},t) \leq 1 \text{ pour presque tout } (\mathbf{x},t) \in \Omega \times (0,T), \\ &u \in L^2\left(0,T; H_0^1(\Omega)\right) \cap \mathcal{C}^0\left([0,T]; L^2(\Omega)\right), \\ &\partial_t u \in L^2\left(0,T; H^{-1}(\Omega)\right), \end{aligned}$$

et telle que

$$\begin{aligned} \int_0^T \langle \partial_t u, \varphi \rangle\, \mathrm{d}t &+ \int_{Q_T} a(u) \nabla u \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_{Q_T} f(u)\, \mathbf{V} \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &+ \int_{Q_T} g(u)\, \mathbf{V} \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t + \int_{Q_T} g'(u)\, \mathbf{V} \cdot \nabla u \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &+ \int_{Q_T} a(u)\, \mathbf{V} \cdot \nabla u \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t = 0, \qquad \forall \varphi \in L^2\left(0,T; H_0^1(\Omega)\right). \end{aligned} \tag{1.32}$$

La notation $\langle \cdot, \cdot \rangle$ représente le crochet de dualité entre $H^{-1}(\Omega)$ et $H_0^1(\Omega)$.

Theorem 1.19. *Sous les hypothèses* (H1) − (H3) *et* (H4a)*, il existe au moins une solution faible classique du système* (1.28)–(1.30) *au sens de la définition 1.18.*

Le deuxième résultat : les solutions faibles dégénérées Afin de définir une notion d'une solution faible adaptée avec l'hypothèse (H4b), nous introduisons les fonctions β et h définies sur $\mathbb{R}$ par :

$$\beta(u) = u^{r-1}, \quad h(u) = \int_0^u \beta(\tau)\, \mathrm{d}\tau, \text{ où } r = \begin{cases} r_1 + 2, & \text{si } r_1 \leq 1, \\ r_1, & \text{si } r_1 > 1. \end{cases}$$

r_1 est la même constante définie dans l'hypothèse (H4b).
Pour un paramètre fixé $\theta \geq 0$, nous considérons les fonctions β_θ et h_θ définies par :

$$\beta_\theta(u) = u^{r-1+\theta}, \quad h_\theta(u) = \int_0^u \beta_\theta(\tau)\, \mathrm{d}\tau.$$

Definition 1.20. Soit $\theta \geq 7r_1 + 6 - r$, sous les hypothèses (H1)–(H3) et (H4b). Nous disons que u est une solution faible dégénérée du système (1.28)–(1.30) si

$$\begin{aligned} &0 \leq u(\mathbf{x}, t) \leq 1 \text{ pour presque tout } (\mathbf{x}, t) \in \Omega \times (0, T), \\ &h_\theta(u) \in L^2\big(0, T; H_0^1(\Omega)\big), \\ &\sqrt{\beta'(u)}\nabla u \in \big(L^2(Q_T)\big)^d, \end{aligned}$$

et telle que la fonction F définie, pour tout $\chi \in L^2\left(0, T; H^1(\Omega)\right)$ par :

$$\begin{aligned} F(u, \chi) = &- \int_{Q_T} h_\theta(u)\, \partial_t \chi \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_\Omega h_\theta(u_0)\, \chi(\mathbf{x}, 0)\, \mathrm{d}\mathbf{x} + \int_{Q_T} a(u)\, \nabla u \cdot \nabla\left(\beta_\theta(u)\, \chi\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &+ \int_{Q_T} a(u)\, \mathbf{V} \cdot \nabla u \beta_\theta(u)\, \chi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_{Q_T} (f(u) - g(u))\, \mathbf{V} \cdot \nabla\left(\beta_\theta(u)\, \chi\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &+ \int_{Q_T} g'(u)\, \mathbf{V} \cdot \nabla u \beta_\theta(u)\, \chi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t, \end{aligned}$$

vérifie

$$F(u, \chi) \leq 0, \quad \forall \chi \in \mathcal{C}^1\left([0, T); H_0^1(\Omega)\right) \text{ avec } \chi(\cdot, T) = 0 \text{ et } \chi \geq 0.$$

De plus, elle vérifie

$$\begin{aligned} &\forall \varepsilon > 0, \exists\, Q^\varepsilon \subset Q_T \text{ tel que mes}\left(Q^\varepsilon\right) < \varepsilon, \text{ et} \\ &F(u, \chi) = 0, \quad \forall \chi \in \mathcal{C}^1\left([0, T); H_0^1(\Omega)\right), \text{supp}\, \chi \subset ([0, T) \times \Omega) \setminus Q^\varepsilon \end{aligned}$$

Theorem 1.21. *Sous les hypothèses* (H1) − (H3) *et* (H4b)*, il existe au moins une solution faible dégénérée du système* (1.28)–(1.30) *au sens de la définition 1.20.*

Le troisième résultat : les solutions faibles dégénérées Afin de définir une notion de solution faible adaptée avec l'hypothèse (H4c), nous introduisons la fonction continue $j_{\theta,\lambda}$ définie par :

$$j_{\theta,\lambda}(u) = \begin{cases} \beta_\theta(u), & \text{si } 0 \leq u \leq u_* \\ \beta_\theta(u_*)\,(1 - u_*)^{1 - \frac{r'}{2} - \lambda}(1 - u)^{\frac{r'}{2} - 1} + \lambda, & \text{si } u \geq u_*. \end{cases}$$

où $r' \geq \max(2, r_2)$ (r_2 est la constante définie dans (H4c)). Nous définissons encore la fonction $J_{\theta,\lambda}$ qui est la primitive de la fonction $j_{\theta,\lambda}$ et notons $j = j_{0,0}$ et $J = J_{0,0}$. Enfin, nous introduisons les fonctions μ et G définies par :

$$\begin{cases} \mu(u) = \beta(u), & 0 \leq u \leq u_* \\ \mu'(u) = (f(u) - g(u))^{-1} g'(u)\, \mu(u), & u_* \leq u < 1, \end{cases} \text{ et } \quad G(u) = \int_0^u \mu(y)\, \mathrm{d}y.$$

Definition 1.22. Soit $\theta \geq 7r_1+6-r$, $\lambda \geq 7r_2+6-\frac{r'}{2}$. Sous les hypothèses (H1)–(H3) et (H4c), nous disons que u est une solution faible dégénérée du système (1.28)–(1.30) si

$$\begin{aligned} &0 \leq u(\mathbf{x},t) \leq 1 \text{ pour presque tout } (\mathbf{x},t) \in \Omega \times (0,T), \\ &J(u) \in L^2\big(0,T;H_0^1(\Omega)\big), \quad \mu'^{\frac{1}{2}}(u)\, a^{\frac{1}{2}}(u)\, \nabla u \in \big(L^2(Q_T)\big)^d, \end{aligned}$$

et telle que la fonction F définie par

$$\begin{aligned} F(u,\chi) = &-\int_{Q_T} J_{\theta,\lambda}(u)\, \partial_t \chi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_\Omega J_{\theta,\lambda}(u_0(\mathbf{x}))\, \chi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} \\ &+ \int_{Q_T} a(u)\, \nabla u \cdot \nabla\left(j_{\theta,\lambda}(u)\,\chi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T} a(u)\, \mathbf{V}\cdot\nabla u\, j_{\theta,\lambda}(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &- \int_{Q_T} \left(f(u)-g(u)\right)\mathbf{V}\cdot\nabla j_{\theta,\lambda}(u)\,\chi\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T} g'(u)\,\mathbf{V}\cdot\nabla u\, j_{\theta,\lambda}(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &- \int_{Q_T} \left(f(u)-g(u)\right)\mathbf{V}\cdot\nabla\chi\, j_{\theta,\lambda}(u)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \end{aligned}$$

vérifie

$$F(u,\chi) \leq 0, \quad \forall\, \chi \in \mathcal{C}^1\left([0,T);H_0^1(\Omega)\right) \text{ avec } \chi(\cdot,T)=0 \text{ et } \chi \geq 0 \tag{1.33}$$

En plus,

$$\begin{aligned} &\forall \varepsilon>0, \exists\, Q^\varepsilon \subset Q_T \text{ telle que mes}\,(Q^\varepsilon) < \varepsilon \text{ et} \\ &F(u,\chi)=0, \quad \forall\, \chi \in \mathcal{C}^1\big([0,T);H_0^1(\Omega)\big),\ \operatorname{supp}\chi \subset ([0,T)\times\Omega)\setminus Q^\varepsilon \end{aligned} \tag{1.34}$$

Theorem 1.23. *Sous les hypothèses* (H1) − (H3) *et* (H4c)*, il existe au moins une solution faible dégénérée du système* (1.28)–(1.30) *au sens de la définition 1.22.*

Liste de publications

Plusieurs publications sont issues de ce travail : un article accepté pour publication dans une revue internationale :

- M. Ibrahim, M. Saad, *On the efficacy of a control volume finite element method for the capture of patterns for a volume-filling chemotaxis model*, accepté pour publication dans **Computers & Mathematics with Applications** (disponible en ligne),

un article soumis :

- C. Cancès, M. Ibrahim, M. Saad, *On a nonlinear control volume finite element method for degenerate anisotropic reaction–convection–diffusion systems modeling the chemotaxis process*,

et un proceeding pour une conférence internationale avec comité de lecture :

- C. Cancès, M. Ibrahim, M. Saad, *A nonlinear CVFE scheme for an anisotropic degenerate nonlinear Keller-Segel model*, **Journal of Mathematics in Industry**, Springer.

2

Pattern formation and an upstream finite-volume scheme for a volume-filling chemotaxis model

Sommaire

2.1 Introduction

Reaction–diffusion systems self-organize variety of spatio-temporal patterns such as propagating circular, spiral, and periodic waves. Those spatio-temporal patterns are observed for the first time in the Belousov-Zhabotinsky reaction system [81].
A reaction–diffusion system is generally described with a set of time-evolving partial differential equations. Each equation consists of a diffusion equation coupled with a reaction term ; the

reaction term usually describes a nonlinear phenomenon observed in nature. Most typical form of reaction–diffusion system is described with a pair of reaction–diffusion equations having an activator variable u and an inhibitor variable v, as follows

$$\begin{cases} \partial_t u = D_u \, \Delta u + f\,(u, v)\,, \\ \partial_t v = D_v \, \Delta v + g\,(u, v)\,, \end{cases} \tag{2.1}$$

where the functions $f\,(u, v)$ and $g\,(u, v)$ denote the reaction kinetics associated with the chemicals u and v. The variables u and v are defined in a two-dimensional bounded domain $(\mathbf{x}, \mathbf{y})$ and in time t with, typically, zeros-flux boundary conditions ; D_u and D_v are the diffusion coefficients of u and v respectively. The system (2.1) with activator and inhibitor variables is a typical model in describing the pattern formation process. Indeed, the mathematician Alan Turing established in his seminal paper [76] in 1952 that, under certain conditions, chemicals can react and diffuse in such a way as to produce inhomogeneous spatial patterns of chemical concentrations. Although a diffusion process generally brings uniform distribution of a substance, the scenario of Turing presents a nonuniform pattern as a stationary state in a reaction–diffusion system.
The idea of Turing is a simple but profound one. He said that, if in the absence of diffusion (effectively $D_u = D_v = 0$), u and v tend to a linearly stable uniform steady state then, under certain conditions called **Turing conditions**, spatially inhomogeneous patterns evolve by *diffusion driven instability* also called *Turing instability* if $D_u \neq D_v$, in more particular when the inhibitor variable v rapidly diffuses more then the activator variable u does ; see, e.g. [60, 61, 58]. This concept was a novel concept because the diffusion was usually considered of having a stabilizing effect.

We can find a wide variety of reaction–diffusion systems in the fields of physics, chemistry and biology [60]. One of the most popular reaction–diffusion system is the chemotaxis model.

Chemotaxis is the feature movement of a cell along a chemical concentration gradient either towards the chemical stimulus, in this case the chemical is called chemoattractant, or away from the chemical stimulus and then the chemical is called chemorepellent. The mathematical analysis of chemotaxis models shows a plenitude of spatial patterns such as the chemotaxis models applied to skin pigmentation patterns [63, 67, 80]− that lead to aggregations of one type of pigment cell into a striped spatial pattern. And other models applied to the aggregation patterns in an epidemic disease [9], tumor growth [19], angiogenesis in tumor progression [14], and many other examples.

Theoretical and mathematical modeling of chemotaxis dates to the pioneering works of Patlak in the 1950s [69] and Keller and Segel in the 1970s [51, 52]. The review article by Horstmann [47] provides a detailed introduction into the mathematics of the Keller–Segel (KS) model for chemotaxis. In its original form, this model consists of four coupled reaction-advection-diffusion equations. These can be reduced under quasi-steady-state assumptions to a model for two unknown functions u and v. The general form of the Keller–Segel model in which we are interested is given by the following system

$$\begin{cases} \partial_t u = \operatorname{div}\left(D\,(u)\,\nabla u - u\,\zeta\,(u, v)\,\nabla v\right) + f\,(u, v), & (\mathbf{x}, t) \in Q_{t_{\mathrm{f}}} = \Omega \times (0, t_{\mathrm{f}}), \\ \partial_t v = D_v \, \Delta v + g\,(u, v), & (\mathbf{x}, t) \in Q_{t_{\mathrm{f}}} = \Omega \times (0, t_{\mathrm{f}}), \end{cases} \tag{2.2}$$

with no-flux boundary conditions on $\Sigma_{t_{\mathrm{f}}} := \partial\Omega \times (0, t_{\mathrm{f}})$,

$$\left(D\,(u)\,\nabla u - u\,\zeta\,(u, v)\,\nabla v\right) \cdot \mathbf{n} = 0, \qquad \nabla v \cdot \mathbf{n} = 0, \tag{2.3}$$

where, $\mathbf{n}$ is the unit outward normal vector at the boundary $\partial\Omega$ of the domain Ω.

In the above model, u denotes the cell density on a given open bounded connected domain $\Omega \subset \mathbb{R}^2$ for a fixed time $t_{\mathrm{f}} > 0$ and v describes the concentration of the chemoattractant. The cell

dynamics derive from population kinetics and movement, the latter comprising a diffusive flux modeling undirected (random) cell migration and an advective flux with velocity dependent on the gradient of the signal, modeling the contribution of chemotaxis. $D(u)$ describes the diffusivity of the cells (sometimes also called motility) while $\zeta(u,v)$ stands to the chemotactic sensitivity. The function $f(u,v)$ describes cell growth and cell death while $g(u,v)$ describes production and degradation of the chemoattractant. A key property of the above equations is their ability to give rise to spatial pattern formation when the chemical signal v acts as an *auto-attractant*, that is, when cells both produce and migrate up gradients of the chemical signal.

The Keller–Segel model (2.2)–(2.3) is studied by many authors, for example we can cite the review articles of Horstmann [46, 47] and the paper of Bendahmane *et al.* [8].

For a particular choice of the diffusivity of the cells $D(u)$, the chemotaxis model (2.2) transforms into a system describing the volume-filling effect. In the volume filling effect, the particles are assumed of having a finite volume and the cells cannot move into regions that are already filled by other cells. Furthermore, in the volume filling approach we consider that the cell jumped into an available space with a given squeezing probability of a cell finding space at its neighboring location see, for example [66].

In [66], Painter and Hillen introduced a squeezing probability of a cell finding space which decreases linearly with the cell density at that site ; indeed, it corresponds to the interpretation that cells are solid blocks and the probability of finding space is proportional to the number of occupants (see, for example [72, 71]). But since the cells are not solid blocks and they are elastic, Wang and Hillen introduced in their paper [78] a nonlinear squeezing probability, which is greater than a linear distribution.

From a numerical point of view, we motion that Bendahmane *et al.* [7] analyzed a finite volume method for the Keller–Segel model (2.2)–(2.3). In this chapter, we adopt the same squeezing probability as in [78] and study the pattern formation for the corresponding volume-filling chemotaxis model using the linear stability analysis and finally we adopt the similar techniques used in [7] to numerically investigate the spatio-temporal patterns in a two-dimensional bounded domain.

The organization of this chapter is as follows. In section 2.2, we define and give the assumptions for the squeezing probability and for the functions f and g that make in evidence global and bounded solutions for the volume-filling chemotaxis model. In section 2.3, we establish Turing conditions for the generation of spatial patterns for the volume-filling chemotaxis model with zeros-flux boundary conditions by performing the standard linear stability analysis and by using a formal asymptotic expansion to linearize the nonlinear diffusion terms. In section 2.4, we take particular functions into the chemotaxis model, apply the pattern formation analysis, and then determine the range of wave numbers for each bifurcation parameter into the chemotaxis model. In section 2.5, we introduce a finite volume scheme to discretize the volume-filling chemotaxis model and to numerically investigate the spatio-temporal patterns in a two-dimensional bounded domain. We end up with some numerical experiments to capture the generation of spatial patterns for the volume-filling chemotaxis model.

2.2 Volume-filling chemotaxis model

In this section, we consider the volume-filling chemotaxis model (2.2)–(2.3) in which we introduce the squeezing probability q of a cell finding space at its neighboring location. q is a non-increasing function with respect to the cell density u and in addition, it is a nonlinear function and equals to zero when the maximum number of cells exceeds a certain number $\overline{u}$. The number $\overline{u}$ represents the total number of cells that can be accommodated at any site.

We take the squeezing probability q introduced in [78]

$$q(u) = \begin{cases} 1 - \left(\frac{u}{\overline{u}}\right)^{\gamma}, & 0 \leq u \leq \overline{u}, \\ 0, & u > \overline{u}, \end{cases} \tag{2.4}$$

with, $\gamma \geq 1$ is a nonnegative exponent representing the rigidity of the cells.
In what follows, the diffusion term and the convection term are given by

$$D(u) = d_1 \left(q(u) - q'(u)\, u\right), \quad \zeta(u, v) = q(u)\, \chi(v),$$

where q is the squeezing probability of a cell finding space into its neighboring location, it is defined in (2.4) and χ is a v-depending continuous function.
Substituting these coefficients into system (2.2), we obtain the final form of the volume-filling chemotaxis model.

We denote by Ω an open bounded domain in $\mathbb{R}^d$, $d = 2, 3$. The final form of the volume-filling chemotaxis model is given by

$$\begin{cases} \partial_t u = \operatorname{div}\left(d_1 \left(q(u) - q'(u)\, u\right) \nabla u - q(u)\, u\, \chi(v)\, \nabla v\right) + f(u, v), & (\mathbf{x}, t) \in Q_{t_{\mathrm{f}}}, \\ \partial_t v = d_2 \Delta v + g(u, v), & (\mathbf{x}, t) \in Q_{t_{\mathrm{f}}}. \end{cases} \tag{2.5}$$

With the following zero flux boundary conditions

$$\begin{aligned} &\left(d_1 \left(q(u) - q'(u)\, u\right) \nabla u\right) \cdot \mathbf{n} - q(u)\, u\, \chi(v)\, \nabla v \cdot \mathbf{n} = 0, \\ &\nabla v \cdot \mathbf{n} = 0, \end{aligned} \tag{2.6}$$

where $\mathbf{n}$ is the unit normal vector at the boundary $\partial\Omega$ and outward to Ω. We assume that the initial conditions are positive

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) \geq 0, \quad v(\mathbf{x}, 0) = v_0(\mathbf{x}) \geq 0, \quad \text{for all } \mathbf{x} \in \Omega.$$

Furthermore, we introduce the assumptions on f, g and q for the global existence of solutions for the system (2.5)

(A1) d_1 and d_2 are two nonnegative constants.

(A2) The chemosensitivity $\chi \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$ such that $\chi(v) > 0$.

(A3) The squeezing probability q is a nonincreasing and concave function lying into $\mathcal{C}^3(\mathbb{R}, \mathbb{R})$ and satisfying the following condition : there exists a critical number $\overline{u}$ such that $q(0) = 1$, $q(\overline{u}) = 0$, $0 < q(u) < 1$ for $u \in (0, \overline{u})$ and $q(u) = 0$ for all $u > \overline{u}$.

(A4) f is a function lying into $\mathcal{C}^2(\mathbb{R} \times \mathbb{R})$ such that $f(0, v) \geq 0$ for all $v \geq 0$. In addition, there exists a constant $0 < u_c < \overline{u}$ such that $f(u_c, v) = 0$ and $f(u, v) < 0$ for all $u > u_c$ and $v \geq 0$.

(A5) g is a function lying into $\mathcal{C}^2(\mathbb{R} \times \mathbb{R})$ such that $g(u, 0) \geq 0$ for all $u \geq 0$. Furthermore, there exists a constant $\overline{v} > 0$ such that $g(u, \overline{v}) < 0$ for all $0 \leq u \leq \overline{u}$.

The existence of global bounded classical solutions of system (2.5) object of the boundary conditions (2.6) has been established in [78] using the maximum principle and the Amman theory [4].

2.3 Pattern formation

By definition, the patterns are the solutions of a reaction–diffusion system which are stable in time and stationary inhomogeneous in space, while the pattern formation in mathematics refers to the process that, by changing a bifurcation parameter, the spatially homogeneous steady states lose stability to spatially inhomogeneous perturbations, and stable inhomogeneous solutions arise.

In this section, we are interested in the pattern formation for system (2.2). We derive here necessary conditions for diffusion-driven instability (also called **Turing instability**) of the steady state and the initiation of spatial pattern for the above-mentioned system. The approach followed here is very routine. We look for the spatial homogeneous steady states by setting the kinetics on the right-hand side of equation (2.5) to be equal to zero,

$$f(u,v)=0, \quad g(u,v)=0. \tag{2.7}$$

We suppose that (u_s, v_s) is the relevant homogeneous steady state of system (2.5), that is mean that (u_s, v_s) is a nonnegative solution of equation (2.7). Since we are concerned with Turing instability, we are interested in linear instability of this steady state that is solely spatially dependent. So, in the absence of any spatial variation, the homogeneous steady state must be linearly stable : we first determine the conditions for this to hold.

With no spatial variation u and v satisfy

$$\partial_t u = f(u,v), \quad \partial_t v = g(u,v). \tag{2.8}$$

2.3.1 Linear stability

Linearizing about the steady state (u_s, v_s) is exactly the same technique used by Murray [60, 61], we set

$$\mathbf{w} = \begin{pmatrix} u-u_s \\ v-v_s \end{pmatrix}, \tag{2.9}$$

and equation (2.8) becomes for $|\mathbf{w}|$ small enough,

$$\partial_t \mathbf{w} = A\mathbf{w}, \quad A = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix}_{u_s,v_s}, \tag{2.10}$$

where A is the stability matrix (i.e. the Jacobian matrix of system (2.8) computed at the steady state (u_s, v_s)). From now on, we take the derivatives of f and g to be evaluated at the steady state (u_s, v_s) unless stated otherwise. We look for solutions in the form

$$\mathbf{w}(t) = \sum_{\lambda} c_\lambda e^{\lambda t}, \tag{2.11}$$

where λ is an eigenvalue of the matrix A. The steady state $\mathbf{w} = 0$ is linearly stable if Re $\lambda < 0$ since in this case the perturbation $\mathbf{w}$ tends to zero as $t \to \infty$. Substituting equation (2.11) into equation (2.10), one can determine the eigenvalues λ as the solutions of the characteristic polynomial

$$|A - \lambda \mathbf{I}_2| = \begin{vmatrix} f_u - \lambda & f_v \\ g_u & g_v - \lambda \end{vmatrix} = 0,$$

where $\mathbf{I}_2$ is the identity matrix in $\mathbb{R}^2$. Hence, the eigenvalues are given by

$$\lambda_1, \lambda_2 = \frac{f_u + g_v \pm \sqrt{(f_u+g_v)^2 - 4f_v g_u}}{2}. \tag{2.12}$$

Therefore, the conditions for which the steady state (u_s, v_s) is linearly stable, i.e. Re $\lambda < 0$ where λ is the eigenvalue of problem (2.10), are guaranteed if

$$\mathrm{tr}(A) = f_u + g_v < 0, \quad \det(A) = f_u\, g_v - f_v\, g_u > 0. \tag{2.13}$$

Whether the kinetics f and g depend on some parameters, then (u_s, v_s) are functions of these parameters, thus the above inequalities impose certain conditions on these parameters.

2.3.2 Formal asymptotic expansion

Now, we consider the full chemotaxis model (2.5). We examine a small perturbation about the homogeneous steady state (u_s, v_s) in the form

$$u = u_s + \varepsilon\tilde{u}\,(\mathbf{x}, t), \quad v = v_s + \varepsilon\tilde{v}\,(\mathbf{x}, t), \tag{2.14}$$

where $\varepsilon \ll 1$ is a small parameter strictly positive.
Since f is a function of class $C^2\,(\mathbb{R}\times\mathbb{R})$, then f can be written using an asymptotic expansion around (u_s, v_s) as follows

$$f\,(u_s + \varepsilon\tilde{u}, v_s + \varepsilon\tilde{v}) = f\,(u_s, v_s) + \varepsilon\tilde{u} f_u\,(u_s, v_s) + \varepsilon\tilde{v} f_v\,(u_s, v_s) + \cdots$$

Performing the same asymptotic expansion for the following functions g, q and χ around (u_s, v_s), u_s, and v_s respectively, and taking into account equation (2.7), then system (2.5) transforms into the following system

$$\begin{cases} \varepsilon\partial_t\tilde{u} = \varepsilon\mathrm{div}\Big(d_1\left(q\,(u_s+\varepsilon\tilde{u}) - q'\,(u_s+\varepsilon\tilde{u})\,(u_s+\varepsilon\tilde{u})\right)\nabla\tilde{u} \\ \qquad\qquad -\varepsilon\,(u_s+\varepsilon\tilde{u})\,\chi\,(v_s+\varepsilon\tilde{v})\,q\,(u_s+\varepsilon\tilde{u})\,\nabla\tilde{v}\Big) \\ \qquad\qquad +\varepsilon\tilde{u} f_u\,(u_s, v_s) + \varepsilon\tilde{v} f_v\,(u_s, v_s) + \cdots \\ \varepsilon\partial_t\tilde{v} = \varepsilon d_2\Delta\tilde{v} + \varepsilon\tilde{u} g_u\,(u_s, v_s) + \varepsilon\tilde{v} g_v\,(u_s, v_s) + \cdots \end{cases} \tag{2.15}$$

Equating first-order terms with respect to ε, neglecting higher-order terms, and dropping the tilde for the sake of brevity, we obtain the following linearized system

$$\begin{cases} \partial_t u = d_1\left(q\,(u_s) - q'\,(u_s)\,u_s\right)\Delta u - u_s q\,(u_s)\,\chi\,(v_s)\,\Delta v + u f_u + v f_v, \\ \partial_t v = d_2\Delta v + u g_u + v g_v. \end{cases} \tag{2.16}$$

An importance of the asymptotic expansion is to transform the nonlinear coefficients of the diffusion and convection terms into constant coefficients which are nothing other than the old coefficients evaluated at the steady state (u_s, v_s).
For simplicity, we denote by $\xi = q\,(u_s) - q'\,(u_s)\,u_s$ and $\psi = -q\,(u_s)\,\chi\,(v_s)\,u_s$ these coefficients. The quantity v_s in $\chi\,(v_s)$ will be often abbreviated for notational convenience unless stated otherwise, i.e. $\chi = \chi\,(v_s)$. Taking into accounts the assumptions for which there exists a nonnegative solution, then one can deduce that $\xi > 0$ and $\psi < 0$.

2.3.3 Turing conditions

In what follows, we assume that the domain is a two dimensional bounded domain. We perform the standard argument (see e.g. [61, chapter II]). We linearize again about the steady state, which with equation (2.9) is $\mathbf{w} = 0$, to get

$$\partial_t\mathbf{w} = A\mathbf{w} + \mathcal{D}\Delta\mathbf{w}, \quad \mathcal{D} = \begin{pmatrix} d_1\xi & \psi \\ 0 & d_2 \end{pmatrix}. \tag{2.17}$$

To solve this system of equations under the Neumann boundary conditions (due to conditions (2.6) mentioned in **section 2.2** and to the asymptotic expansion), we define the time-independent solution $\mathbf{W}(\mathbf{x})$ of the spatial eigenvalue problem defined by

$$-\Delta \mathbf{W} = k^2 \mathbf{W} \quad \text{in } \Omega, \qquad \nabla \mathbf{W} \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega, \tag{2.18}$$

where k is the eigenvalue. For example, assume that the domain is one-dimensional, say, $0 \leq \mathbf{x} \leq L$. Then, system (2.18) is summarized by the following equation

$$\mathbf{W}''(\mathbf{x}) + k^2 \mathbf{W}(\mathbf{x}) = 0, \quad \mathbf{W}_r'(0) = \mathbf{W}_l'(L) = 0, \quad \forall\, 0 \leq \mathbf{x} \leq L.$$

It easy to see that $\mathbf{W}(\mathbf{x}) = \sum_{n\in\mathbb{Z}} c_n \cos(n\pi \mathbf{x}/L)$ is a nontrivial solution where c_n is a constant different from zero. The eigenvalue in this case is $k = n\pi/L$, it is also called the wavenumber. From now on, we shall refer to k in this context as the wavenumber. With finite domains, there exists a discrete set of possible wave numbers since n is an integer.

Since the problem is linear and by the superposition principle, the linear analysis consists of looking for solutions $\mathbf{W}(\mathbf{x}, t)$ to system (2.17) in the following form

$$\mathbf{W}(\mathbf{x}, t) = \sum_k c_k e^{\lambda t} \mathbf{W}_k(\mathbf{x}), \tag{2.19}$$

where for every wavenumber k, $\mathbf{W}_k$ is the eigenfunction of problem (2.18), c_k are constants determined by a Fourier expansion of the initial conditions in terms of $\mathbf{W}_k(\mathbf{x})$, λ is the eigenvalue which determines temporal growth and it depends on the wavenumber k.

Substituting form (2.19) into equation (2.17) with (2.18), and canceling $e^{\lambda t}$, we get, for each wavenumber k

$$\lambda \mathbf{W}_k = A\mathbf{W}_k - \mathcal{D}k^2 \mathbf{W}_k.$$

We take the nontrivial solutions for $\mathbf{W}_k$ so the λ are nothing other then the eigenvalues of the matrix $A - \mathcal{D}k^2$, which are determined by the roots of the characteristic polynomial

$$\left|\lambda \mathbf{I}_2 - A + \mathcal{D}k^2\right| = 0, \quad \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Consequently, the dispersion relation associated with system (2.16) is determined by the following equation

$$\lambda^2 + a\left(k^2\right)\lambda + b\left(k^2\right) = 0, \tag{2.20}$$

where $a = a\left(k^2\right)$ and $b = b\left(k^2\right)$ are two functions given by

$$\begin{aligned} a\left(k^2\right) &= (\xi d_1 + d_2)\, k^2 - (f_u + g_v), \\ b\left(k^2\right) &= \xi d_1 d_2 k^4 + (\psi g_u - d_2 f_u - \xi d_1 g_v)\, k^2 + f_u\, g_v - f_v\, g_u. \end{aligned}$$

The homogeneous steady state (u_s, v_s) is linearly stable if both solutions of equation (2.20) have $\operatorname{Re}\lambda < 0$ in the absence of any spatial effects that is equivalent to $k = 0$. We have already imposed the constraints that the steady state is stable in the absence of any spatial effects ; that is, $\operatorname{Re}\lambda < 0$. The quadratic (2.20) in this case with the requirement that $\operatorname{Re}\lambda < 0$ gives conditions (2.13). Now, for the steady state to be unstable to spatial perturbations, we require that $\operatorname{Re}\lambda > 0$ for some $k \neq 0$. We remark that the sum of the eigenvalues $\left(= -a\left(k^2\right)\right)$ is strictly negative since the function $a\left(k^2\right)$ is strictly positive due to condition (2.13). Thus, the instability can only happen if $b\left(k^2\right)$ becomes negative for some $k \neq 0$ i.e. the dispersion relation (2.20) has two roots with opposite signs. Carrying the function $b\left(k^2\right)$, then the last condition requires that

$$\xi d_1 d_2 k^4 + (\psi g_u - d_2 f_u - \xi d_1 g_v)\, k^2 + f_u g_v - f_v g_u < 0. \tag{2.21}$$

Since we required that the determinant $\det(A) = f_u\, g_v - f_v\, g_u > 0$ from condition (2.13), the only possibility for inequality (2.21) to hold, requires that the sum of the roots is positive

$$\psi g_u - d_2 f_u - \xi d_1 g_v < 0. \tag{2.22}$$

Referring to condition (2.21), and since $\xi d_1 d_2$ is positive, it is necessary that the discriminant of $b\left(k^2\right) = 0$ to be positive. That is

$$\left(\psi g_u - d_2 f_u - \xi d_1 g_v\right)^2 - 4\xi d_1 d_2 \left(f_u g_v - f_v g_u\right) > 0. \tag{2.23}$$

Applying equation (2.22), we obtain from this inequality that

$$\psi g_u - d_2 f_u - \xi d_1 g_v < -2\sqrt{\xi d_1 d_2 \left(f_u g_v - f_v g_u\right)}.$$

To recap, we have now obtain necessary conditions for the generation of spatial patterns by two-species reaction–diffusion mechanisms of the form (2.5)–(2.6) which correspond to the volume-filling chemotaxis model. For simplicity, we reproduce the conditions here :

$$\begin{aligned} &f_u + g_v < 0, \quad f_u\, g_v - f_v\, g_u > 0, \quad \psi g_u - d_2 f_u - \xi d_1 g_v < 0, \\ &\left(\psi g_u - d_2 f_u - \xi d_1 g_v\right)^2 - 4\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right) > 0. \end{aligned} \tag{2.24}$$

Remembering that all these derivatives are evaluated at the steady state (u_s, v_s).

2.3.4 Bifurcation

Here, we are interested in the mathematical study of changes in the phase portrait for a dynamical system. Such changes called bifurcations help us to determine the parameters for which we can expect pattern formation. System (2.5) depends on many parameters as χ and others involved in the kinetic functions. As well these parameters play a significant role in the production of bifurcations. Once the bifurcation occurs (by changing the value of the bifurcation parameter), we get spatial patterns under the Turing conditions.

The strength of the chemotactic sensitivity χ plays a crucial role in pattern formation. Generally, there exists a critical value χ_c such that there is no pattern formation if χ is below this critical value χ_c, while pattern formation can be expected if χ is somewhere else.

To determine explicitly this critical value, we fix all the parameters in system (2.5) except for the chemotactic sensitivity χ. Effectively, we know from the previous analysis that the bifurcation occurs when $b_{\min} = b\left(k^2_{\min}\right) = 0$, which is equivalent to

$$\psi g_u - d_2 f_u - \xi d_1 g_v = -2\sqrt{\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)}. \tag{2.25}$$

Substituting $\psi = -u_s q\left(u_s\right) \chi\left(v_s\right)$ into equation (2.25), then the critical chemosensitivity χ_c is given in the following expression

$$\chi_c = \frac{2\sqrt{\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)} - d_2 f_u - \xi d_1 g_v}{g_u u_s q\left(u_s\right)}. \tag{2.26}$$

For this bifurcation value, the associated critical wavenumber is given by $b'\left(k_c^2\right) = 0$, that is

$$k_c^2 = \frac{d_2 f_u + \xi d_1 g_v + u_s q\left(u_s\right) \chi_c\, g_u}{2\xi d_1 d_2}. \tag{2.27}$$

Whenever $b\left(k^2\right)$ is negative, the equation (2.20) has a solution λ which is positive for the same range of wave numbers that make $b\left(k^2\right) < 0$. When $\chi > \chi_c$, the range of unstable wave numbers $k_1^2 < k^2 < k_2^2$ is obtained from the zeros k_1^2 et k_2^2 of equation $b\left(k^2\right) = 0$, and it is defined as

$$\begin{aligned} k_1^2 &= \frac{C - \sqrt{C^2 - 4\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)}}{2\xi d_1 d_2} < k^2 \\ &< k_2^2 = \frac{C + \sqrt{C^2 - 4\xi d_1 d_2 \left(f_u\, g_v - f_v\, g_u\right)}}{2\xi d_1 d_2}, \end{aligned} \tag{2.28}$$

where $C = d_2 f_u + \xi d_1 g_v + u_s q\left(u_s\right)\chi > 0$ is nothing else the symmetric of the coefficient of k^2 in equation $b\left(k^2\right) = 0$.
From all the foregoing analysis, under conditions (2.24) and if there exists a range of wave numbers k satisfying condition (2.28), then the corresponding spatial eigenfunctions are linearly unstable and the generation of spatial patterns can be expected.

2.4 Analysis for a nonlinear density

Hereinafter, we introduce the probability function $q\left(u\right)$ which reflects the elastic property of particles

$$q\left(u\right) = 1 - \left(\frac{u}{\bar{u}}\right)^{\gamma}, \quad \gamma \geq 1,\ 0 \leq u \leq \bar{u},$$

where $\bar{u}$ is the total number of cells that can be accommodated at any site. We assume that the cell density and the chemoattractant kinetics are respectively given by

$$f\left(u, v\right) = \mu u\left(1 - \frac{u}{u_c}\right), \quad g\left(u, v\right) = \alpha u - \delta v.$$

This choice is the best choice for the cell kinetics where it represents the logistic growth with carrying capacity $0 < u_c \leq \bar{u}$ and $\mu > 0$; while the choice for chemoattractant kinetics is extensively used in literature (see, e.g. [48]) which represents the chemoattractant growth with rate α and the death with rate δ due to dilution.
Substituting the above mentioned functions into system (2.5) and assuming that the function χ is a constant function ; then, we get the following system :

$$\begin{cases} \partial_t u = \operatorname{div}\left(D\left(u\right)\nabla u - \chi\varphi\left(u\right)\nabla v\right) + \mu u\left(1 - \dfrac{u}{u_c}\right), & \text{in } Q_{t_{\mathrm{f}}} \\ \partial_t v = d_2\Delta v + \alpha u - \beta v, & \text{in } Q_{t_{\mathrm{f}}} \\ \left(D\left(u\right)\nabla u - \chi\varphi\left(u\right)\nabla v\right)\cdot \mathbf{n} = 0, \quad \nabla v \cdot \mathbf{n} = 0, & \text{in } \partial\Omega \times \left(0, t_{\mathrm{f}}\right) \end{cases} \tag{2.29}$$

where, $\mathbf{n}$ is the unit outward normal vector at the boundary $\partial\Omega$ of the domain Ω and the functions $D\left(u\right)$ et $\varphi\left(u\right)$ are given by

$$D\left(u\right) = d_1\left(1 + \left(\gamma - 1\right)\left(\frac{u}{\bar{u}}\right)^{\gamma}\right), \quad \varphi\left(u\right) = u\left(1 - \left(\frac{u}{\bar{u}}\right)^{\gamma}\right).$$

The homogeneous steady states of system (2.29) are $\left(0, 0\right)$ and $\left(u_c, \alpha u_c\,/\beta\right)$. In addition, by linearization, one can determine that the steady state $\left(0, 0\right)$ is a saddle point and consequently unstable, while the homogeneous steady state $\left(u_c, \alpha u_c\,/\beta\right)$ is stable to the corresponding homogeneous system of (2.29). Indeed, computing the matrix stability at the steady state $\left(u_s, v_s\right) = \left(u_c, \alpha u_c\,/\beta\right)$, one can determine the eigenvalues corresponding to that matrix and find two negative eigenvalues $-\alpha$ et $-\beta$, thus conditions (2.13) are verified and we focus on this steady state to study pattern

formation for the system (2.29).
Let us begin by the pattern formation analysis for the system (2.29) associated to functions q, f, and g ; we apply an asymptotic expansion around the steady state $(u_c, \alpha u_c/\beta)$, one get

$$\begin{cases} \partial_t u = \operatorname{div}\left(D(u_c)\nabla u - \chi\varphi(u_c)\nabla v\right) - \mu u, & \text{in } Q_{t_f}, \\ \partial_t v = d_2\Delta v + \alpha u - \beta v. & \text{in } Q_{t_f}. \end{cases} \tag{2.30}$$

Next, we follow the same analysis as in the previous section to end up with the following relation dispersion

$$\lambda^2 + a\left(k^2\right)\lambda + b\left(k^2\right) = 0, \tag{2.31}$$

where $a\left(k^2\right)$ and $b\left(k^2\right)$ are two functions given by

$$\begin{aligned} a\left(k^2\right) &= \left(D(u_c) + d_2\right)k^2 + (\mu + \beta), \\ b\left(k^2\right) &= D(u_c)\,d_2k^4 + \left(\mu d_2 + \beta D(u_c) - \chi\alpha\varphi(u_c)\right)k^2 + \mu\beta. \end{aligned} \tag{2.32}$$

Thereby, taking into account the stability matrix corresponding to the steady state $(u_c, \alpha u_c/\beta)$, then conditions (2.24) are verified if and only if

$$\mu d_2 + \beta D(u_c) - \chi\alpha\varphi(u_c) < -2\sqrt{d_2\mu\beta D(u_c)}. \tag{2.33}$$

The condition (2.33) is a necessary condition for the pattern formation of the system (2.30) with the specific functions q, f, and g. It is also a sufficient condition for an infinite domain while for a finite domain, there exists a range of discrete wave numbers depending in part on the boundary conditions, and for which we can expect the generation of spatial patterns. Furthermore, we have to study the influence of the bifurcation parameters to determine that range of wave numbers.

In the remainder of this section, we will investigate the influence of each of the parameters, on which depends the system (2.29), on the pattern formation for system (2.29) . The main parameters we are interested in are : the squeezing exponent γ, the chemosensitivity χ, the growth rate α, and the death rate β of the chemoattractant. For each of these parameters, we determine the range of wave numbers for which there exists spatial pattern formation for the system (2.29).

2.4.1 Bifurcation with chemotactic sensitivity χ

From the previous section, we can think about the chemotactic sensitivity χ as a bifurcation parameter, we determine the bifurcation value χ_c to obtain

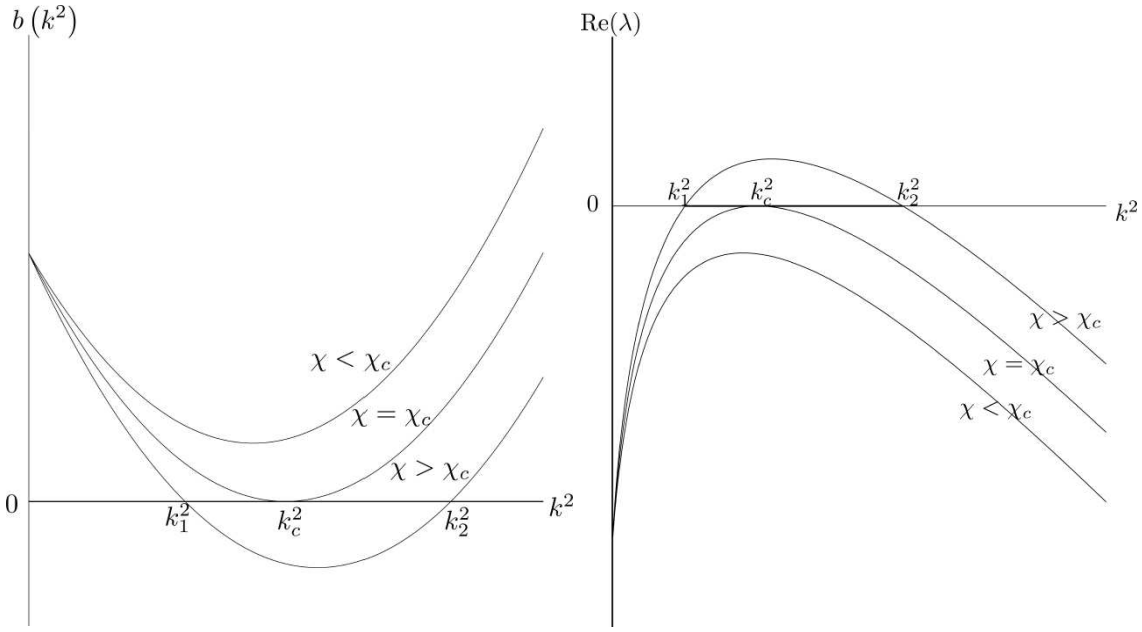
$$\chi_c = \frac{2\sqrt{d_2\mu\beta D(u_c)} + \mu d_2 + \beta D(u_c)}{\alpha\varphi(u_c)}. \tag{2.34}$$

The corresponding critical wave number k_c is determined from equation (2.27) by

$$k_c^2 = \frac{-\mu d_2 - \beta D(u_c) + \chi_c\alpha\varphi(u_c)}{2d_2D(u_c)}. \tag{2.35}$$

Whenever $b\left(k^2\right)$ is negative, equation (2.31) has a solution λ which is positive for the same range of wave numbers that make $b = b\left(k^2\right)$ negative. Furthermore, the range of unstable wave numbers $k_1^2 < k^2 < k_2^2$ is obtained from the zeros k_1^2 and k_2^2 of equation $b\left(k^2\right) = 0$, it is defined as

$$k_1^2 = \frac{S - \sqrt{S^2 - 4\mu d_2\beta D(u_c)}}{2\xi d_1 d_2} < k^2 < \frac{S + \sqrt{S^2 - 4\mu d_2\beta D(u_c)}}{2\xi d_1 d_2} = k_2^2,$$



(a) (b)

FIGURE 2.1 – (a) Plot of $b\left(k^2\right)$ as a function of k^2 defined by equation (2.32). When the chemosensitivity strength χ increases beyond the critical value χ_c, $b\left(k^2\right)$ becomes negative for a finite range of $k^2 > 0$.
(b) Plot of the real part of the eigenvalue $\lambda\left(k^2\right)$ as a function of k^2 defined by equation (2.31). When $\chi > \chi_c$, there is a range of wave numbers $k_1^2 < k^2 < k_2^2$ for which the homogeneous steady state is linearly unstable.

where $S = -\mu d_2 - \beta D\left(u_c\right) + \chi\alpha\varphi\left(u_c\right)$.
Figure 2.1a represents the variation of $b\left(k^2\right)$ as a function of k^2 and for various values of χ. Note that the critical value is (k_c, χ_c) such that $b\left(k^2\right) > 0$ if $\chi < \chi_c$, for all $k^2 > 0$, nevertheless, $b\left(k^2\right) < 0$ if $\chi > \chi_c$ and for a range of unstable wave numbers $k_1^2 < k^2 < k_2^2$.
Figure 2.1b represents the variation of the dispersion relation $\lambda\left(k^2\right)$ as a function of k^2 and for various values of χ. This graph is extremely informative in that it immediately says which eigenfunctions are linearly unstable and grow exponentially with time. Indeed, if we consider the solution $\mathbf{W}$ given by equation (2.19), the dominant contributions as t increases are those modes for which $\mathrm{Re}\,\lambda\left(k^2\right) > 0$ since all other modes tend to zero exponentially. Thus, from Fig. 2.1, we determine the range $k_1^2 < k^2 < k_2^2$, where $b\left(k^2\right) < 0$, and hence $\mathrm{Re}\,\lambda\left(k^2\right) > 0$, and the eigenfunctions are unstable.

Relationship between the critical value χ_c and the exponent γ

Herein, we are interested in the relationship between the squeezing exponent γ and the critical value χ_c. A such relation is important to know the influence of the parameter γ on the dynamics of system (2.29). Indeed, the greater the critical value is smaller, the more pattern formation will be fast, here we show that the critical value χ_c as a function of the exponent γ, is nonincreasing. Hence γ plays an important role in the pattern formation for system (2.29).

We denote by M the function defined by

$$M\left(\gamma\right) = \frac{1}{1 - \left(\frac{u_c}{\bar{u}}\right)^{\gamma}},$$

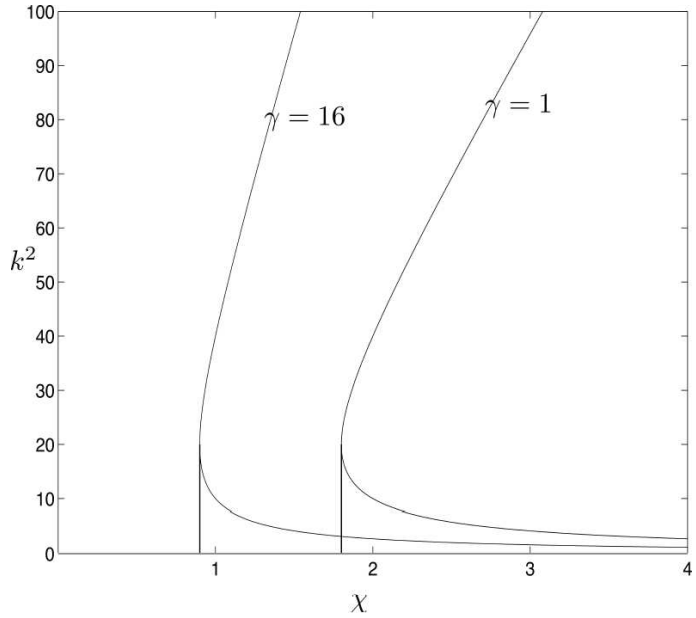


FIGURE 2.2 – Graphic representing the influence of the exponent parameter γ on the critical value χ_c. When γ increases, χ_c decreases in value. The other parameters of system (2.30) are fixed and chosen as $d_1 = 0.1$, $d_2 = 1.0$, $u_c = 2.0$, $\bar{u} = 4.0$, $\mu = 4.0$, $\alpha = 5.0$, $\beta = 10.0$.

then M is nonincreasing with respect to γ, since $u_c/\bar{u} < 1$.
Now, writing the critical value χ_c as a function of $M(\gamma)$, on can easily verify that χ_c is decreasing as a function of the exponent γ . Thereby the pattern formation is easier to form when γ is great.
Figure 2.2 represents for different values of γ, the variation of the critical value $\left(\chi_c, k_c^2\right)$ with respect to γ. Note that the greater the value of γ is bigger, the more value χ_c decreases while remaining minus the value 0.9 and for the fixed parameters.

2.4.2 Bifurcation with growth rate α

In this part, we consider the growth rate α as the bifurcation parameter. We apply the same method used in section 2.4.1, i.e. we fix all the parameters in system (2.29) except the parameter α. We want to study the influence of the dynamical parameter α on the pattern formation for system (2.29).
We compute the critical value α_c for the growth rate α, we obtain

$$\alpha_c = \frac{2\sqrt{d_2\mu\beta D(u_c)} + \mu d_2 + \beta D(u_c)}{\chi\varphi(u_c)} \tag{2.36}$$

By comparing this value with the critical value χ_c in equation (2.34), one can note that there is no unstable modes if α is below the critical value α_c, whereas unstable modes are possible when α is beyond this critical value α_c.
Figure 2.3 represents the plot of the relation dispersion as a function of k^2 and for various values of α, we remark that we have consistent results to those in figure 2.1a concerning the existence of a range of wave numbers $k_1^2 < k^2 < k_2^2$ for which the eigenfunctions are unstable. Consequently, the pattern formation for system (2.29) can be expected by increasing the value of the dynamical parameter α above the critical value α_c.

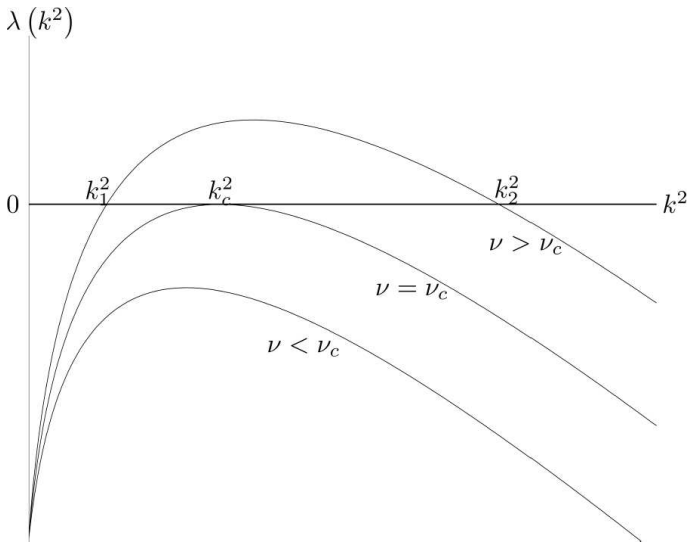


FIGURE 2.3 – Plot of the relation dispersion (2.31) as the growth rate parameter α passes through the critical value α_c. The parameters of the system (2.30) are chosen as $d_1 = 0.1$, $d_2 = 1.0$, $u_c = 2.0$, $\bar{u} = 4.0$, $\mu = 4.0$, $\alpha = 5.0$, $\beta = 10.0$, $\chi = 0.5$

2.4.3 Bifurcation with death rate β

We are interested now by the study of the influence of the death rate β on the pattern formation for system (2.29). We perform the similar linear stability as we did in the previous sections. We compute the critical value β_c for the death rate β, we obtain

$$\chi\alpha\varphi\left(u_c\right) = \mu d_2 + \beta_c D\left(u_c\right) + 2\sqrt{d_2\mu\beta_c D\left(u_c\right)} = \left(\sqrt{d_2\mu} + \sqrt{\beta_c D\left(u_c\right)}\right)^2,$$

that is

$$\beta_c = \frac{\left(\sqrt{\chi\alpha\varphi\left(u_c\right)} - \sqrt{d_2\mu}\right)^2}{D\left(u_c\right)}. \tag{2.37}$$

Taking into account the condition (2.33), one can note that there do not exist unstable modes when $\beta > \beta_c$, whereas unstable modes are possible when β is below the critical value β_c.
Fig. 2.4 represents the plot of the relation dispersion as a function of k^2 and for several values of β. We observe that the plot of the relation dispersion as β passes through the critical value β_c crosses the horizontal axis of k^2 and as result, there exists a range of unstable wave numbers $k_1^2 < k^2 < k_2^2$. Hence, the pattern formation can be expected by decreasing the value of the dynamical parameter β below the critical value β_c.

2.5 Finite volume approximation

In this section, we propose a numerical scheme to investigate the generation of spatial patterns for the volume-filling chemotaxis model (2.29) in a two dimensional space. We use the finite volume method to discretize the diffusion terms over an unstructured mesh namely admissible mesh (see [31]). Whereas, the convective term is discretized using a classical upwind finite volume scheme. The resulted scheme ensures the validity of the discrete maximum principle under some conditions on the mesh. Maintaining the discrete maximum principle is one of the important features of the finite volume scheme. The local conservativity of the numerical fluxes is an another feature of the finite volume scheme, i.e. the numerical flux is conserved from one discretization cell to its neighbor.

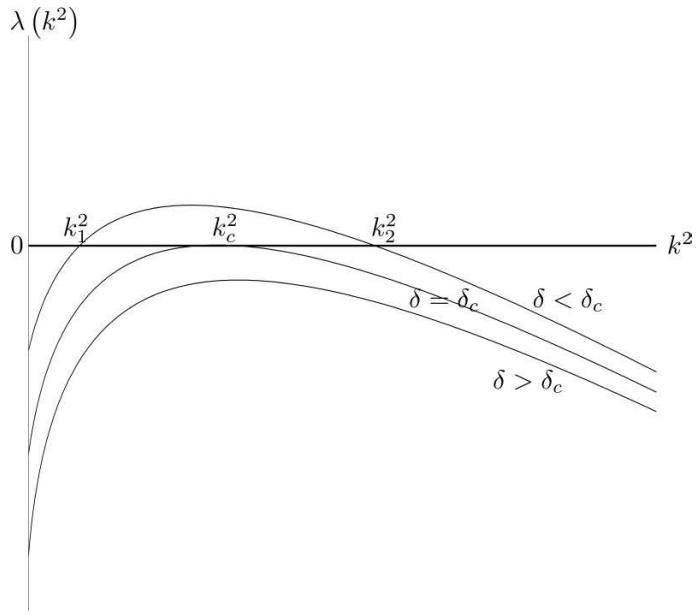


FIGURE 2.4 – Plot of the relation dispersion (2.31) as the death rate parameter β passes through the critical value β_c. For the other parameters of system (2.30), we fix $d_1 = 0.1$, $d_2 = 1.0$, $u_c = 2.0$, $\bar{u} = 4.0$, $\mu = 10.0$, $\alpha = 20$, $\chi = 0.5$, and we find that the relation dispersion passes through the bifurcation value β_c.

2.5.1 Space-time discretization and discrete functions

In order to discretize system (2.29), we introduce some definitions and notations. We assume that $\Omega \subset \mathbb{R}^2$, is an open bounded polygonal connected domain with boundary $\partial\Omega$.

Space discretization

Definition 2.1 (Admissible mesh)**.** An admissible finite volume mesh of Ω is a triplet $(\mathcal{T}, \mathcal{E}, \mathcal{P})$, where $\mathcal{T}$ is a finite family of disjoint open polygonal convex subsets of Ω called control volumes, $\mathcal{E}$ is a finite family of subsets of $\overline{\Omega}$ contained in straights of $\mathbb{R}^2$ with strictly positive one-dimensional measure, called the edges of the control volumes, and $\mathcal{P} = \{\mathbf{x}_K, K \in \mathcal{T}\}$ is a finite family of points of Ω, called the centers of the control volumes.
The triplet $(\mathcal{T}, \mathcal{E}, \mathcal{P})$ satisfies the following properties :

1. The closure of the union of all the control volumes is $\overline{\Omega}$, i.e. $\overline{\Omega} = \bigcup_{K\in\mathcal{T}} \overline{K}$.
2. For any $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \overline{K}\backslash K = \bigcup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$. Furthermore, $\mathcal{E} = \bigcup_{K\in\mathcal{T}} \mathcal{E}_K$.
3. For any $(K, L) \in \mathcal{T}^2$ with $K \neq L$, either the length of $\overline{K} \cap \overline{L}$ is 0 or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$, which then is denoted by σ_{KL} i.e. the interface between K and L.
4. The family $\mathcal{P} = (\mathbf{x}_K)_{K\in\mathcal{T}}$ is such that $\mathbf{x}_K \in \overline{K}$ (for all $K \in \mathcal{T}$) and, if $\sigma = \sigma_{KL}$, we assumed that $\mathbf{x}_K \neq \mathbf{x}_L$, and the straight line $\mathscr{D}_{KL}$ going through $\mathbf{x}_K$ and $\mathbf{x}_L$ is orthogonal to σ_{KL}.
5. For any $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega$, let K the control volume such that $\sigma \in \mathcal{E}_K$. If $\mathbf{x}_K \notin \sigma$, let $\mathscr{D}_{K,\sigma}$ be the straight line going through $\mathbf{x}_K$ and orthogonal to σ, then the condition $\mathscr{D}_{K,\sigma} \cap \sigma \neq \emptyset$ is assumed ; let $\mathbf{y}_\sigma = \mathscr{D}_{K,\sigma} \cap \sigma$.

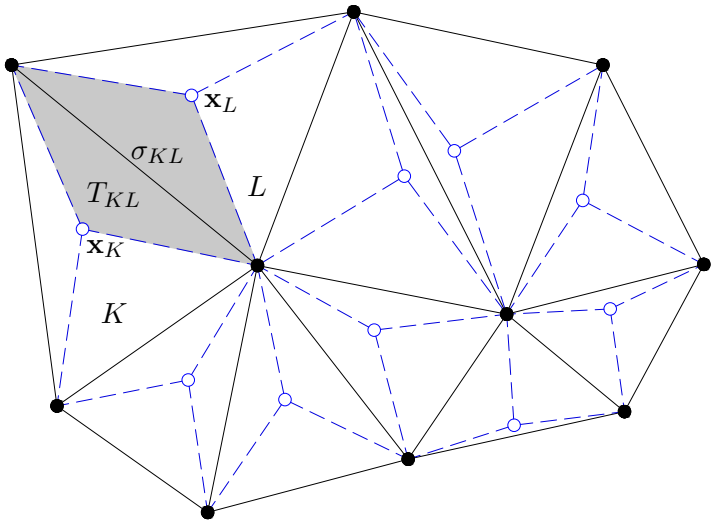


FIGURE 2.5 – Finite volume mesh $\mathcal{T}$: control volumes, centers and diamonds.

In the sequel, we use the following notations. The size of the mesh $\mathcal{T}$ is defined by :

$$h = \text{size}\,(\mathcal{T}) = \sup\{\text{diam}\,(K)\,, K \in \mathcal{T}\},$$

where $\text{diam}(K)$ represents the least upper bound of the set of all distances between pairs of vertices of every control volume $K \in \mathcal{T}$.
For any $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}$, we denote by $|K|$ the 2-dimensional Lebesgue measure of K (it is the area of K in the 2-dimensional case) and by $|\sigma|$ the length of the interface σ.
The set of neighbors of K is denoted by $\mathcal{N}(K)$, i.e. $\mathcal{N}(K) = \{L \in \mathcal{T}; \exists \sigma \in \mathcal{E}_K, \overline{\sigma} = \overline{K} \cap \overline{L}\}$; a generic neighbor of K is often denoted by L, furthermore we denote by η_{KL} and d_{KL} the unit normal vector to σ_{KL} outward to K and the distance $|\mathbf{x}_K - \mathbf{x}_L|$ respectively.

Time discretization

The time discretization is considered to be uniform (we do not impose any restriction on the time step), it is given by the sequence of discrete time $t_n = n\Delta t$ for every $n \in \mathbb{N}$ with a fixed time step Δt such that there exists an integer $N \in \mathbb{N}^*$ verifying $t_{N+1} = (N+1)\,\Delta t = t_{\text{f}}$.

Discrete space $H_{\mathcal{T}}$

Let us define the discrete finite volume space $H_{\mathcal{T}}$ of piecewise constant functions on the mesh $\mathcal{T}$ by

$$H_{\mathcal{T}} = \{\phi : \Omega \longrightarrow \overline{\mathbb{R}};\ \phi_{|_K} \in \overline{\mathbb{R}} \text{ is constant},\ \forall K \in \mathcal{T}\}.$$

Every function $u_{\mathcal{T}} \in H_{\mathcal{T}}$ is characterized by its numerical values $(u_K)_{K\in\mathcal{T}}$ such that $u_{\mathcal{T}|_K} = u_K$ for every $K \in \mathcal{T}$. In other words, given a vector $(u_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}$; then, there exists a unique discrete function $u_{\mathcal{T}} \in H_{\mathcal{T}}$ such that

$$u_{\mathcal{T}}(\mathbf{x}) = u_{\mathcal{T}}(\mathbf{x}_K) = u_K, \quad \forall \mathbf{x} \in K,\ \forall K \in \mathcal{T}.$$

The discrete space $H_{\mathcal{T}}$ is a linear subspace of $L^2(\Omega)$, the usual inner scalar product becomes

$$(u_{\mathcal{T}}, v_{\mathcal{T}})_{L^2(\Omega)} = \int_\Omega u_{\mathcal{T}}(\mathbf{x})\, v_{\mathcal{T}}(\mathbf{x})\, \mathrm{d}\mathbf{x} = \sum_{K\in\mathcal{T}} |K|\, u_K v_K, \quad \forall u_{\mathcal{T}}, v_{\mathcal{T}} \in H_{\mathcal{T}}.$$

Therefore, the associated norm is given by

$$\|u_{\mathcal{T}}\|_{L^2(\Omega)} = \left(\sum_{K \in \mathcal{T}} |K| \, |u_K|^2 \right)^{1/2}.$$

We define a discrete equivalent to the semi-norm on $H^1(\Omega)$, $|u|_{1,\Omega} = \left(\int_\Omega |\nabla u|^2 \right)^{1/2}$, by

$$|u_{\mathcal{T}}|_{1,\mathcal{T}} = \left(\sum_{\sigma_{KL} \in \mathcal{E}} \frac{|\sigma_{KL}|}{d_{KL}} |u_L - u_K|^2 \right)^{1/2},$$

for all $u_{\mathcal{T}} \in H_{\mathcal{T}}$.

Definition 2.2. Let $\sigma \in \mathcal{E}$ be an interface for a control volume $K \in \mathcal{T}$. A "diamond" T_σ constructed from the neighbor centers $\mathbf{x}_K$, $\mathbf{x}_L$ and the interface σ_{KL} is defined by

$$\begin{cases} T_{KL} = \bigcup_{V \in \{K,L\}} \{c\mathbf{x}_V + (1-c)\,\mathbf{y},\ c \in [0,1[,\ \mathbf{y} \in \sigma_{KL}\}, & \text{if } \sigma = \sigma_{KL} \in \mathcal{E} \backslash \partial\Omega, \\ T_{K\sigma} = \{c\mathbf{x}_K + (1-c)\,\mathbf{y},\ c \in [0,1[,\ \mathbf{y} \in \sigma_{KL}\}, & \text{if } \sigma \in \partial K \cap \partial\Omega. \end{cases}$$

The 2-dimensional Lebesgue measure of an interior diamond T_{KL} is $|T_{KL}| = \frac{|\sigma_{KL}| d_{KL}}{2}$.

Using definition 2.2, we define the discrete gradient over every "diamond" T_σ. Specifically, we have the following definition.

Definition 2.3 (Discrete gradient)**.** The discrete gradient $\nabla_{\mathcal{T}}$ is a correspondence that maps a function $u_{\mathcal{T}} \in H_{\mathcal{T}}$ into a piecewise constant function over the "diamonds" such that

$$\nabla_{\mathcal{T}} u_{\mathcal{T}|_{T_\sigma}} = \begin{cases} 2\dfrac{u_L - u_K}{d_{KL}} \eta_{KL}, & \text{if } \sigma = \sigma_{KL} \in \mathcal{E} \backslash \partial\Omega, \\ 0, & \text{if } \sigma \in \partial K \cap \partial\Omega. \end{cases}$$

Remark 1. From the definition 2.3 of the discrete gradient, one can deduce that

$$\|\nabla_{\mathcal{T}} u_{\mathcal{T}}\|_{L^2(\Omega)} = \left(2 \sum_{\sigma_{KL} \in \mathcal{E}} \frac{|\sigma_{KL}|}{d_{KL}} |u_L - u_K|^2 \right)^{1/2} = \sqrt{2}\, |u_{\mathcal{T}}|_{1,\mathcal{T}}.$$

Definition 2.4 (Discrete functions)**.** Using the values of (u_K^n, v_K^n), $\forall K \in \mathcal{T}$ and $n \in \{0, \cdots, N+1\}$, we determine the approximate finite volume solutions as piecewise constant on the control volumes and piecewise in time such that

$$\begin{cases} (u_{\mathcal{T},\Delta t}(\mathbf{x},0), v_{\mathcal{T},\Delta t}(\mathbf{x},0)) = \left(u_K^0, v_K^0\right), & \forall \mathbf{x} \in \overset{\circ}{K},\ K \in \mathcal{T}, \\ (u_{\mathcal{T},\Delta t}(\mathbf{x},t), v_{\mathcal{T},\Delta t}(\mathbf{x},t)) = \left(u_K^{n+1}, v_K^{n+1}\right), & \forall \mathbf{x} \in \overset{\circ}{K},\ K \in \mathcal{T},\ \forall t \in (t_n, t_{n+1}], \end{cases}$$

where the quantity u_K^0 (resp. v_K^0) represents the mean value of the function u_0 (resp. of the function v_0). The discrete time-space of these functions is denoted by $H_{\mathcal{T},\Delta t}$.

In what follows, we use the Lipschitz nondecreasing function $A : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$A(u) = \int_0^u \Gamma(s)\, \mathrm{d}s, \tag{2.38}$$

where $\Gamma : \mathbb{R} \longrightarrow \mathbb{R}$ is the function defined by $\Gamma(u) = 1 + (\gamma - 1)\left(\frac{u}{\bar{u}}\right)^\gamma$, so that $D(u) = d_1 \Gamma(u)$. The function A is nonlinear, we denote by $A(u_{\mathcal{T},\Delta t})$ the corresponding piecewise constant reconstruction in $H_{\mathcal{T},\Delta t}$.

2.5.2 Finite volume scheme for system (2.29)

Once the space-time discretization of the domain has been built, we can discretize the volume filling chemotaxis model (2.29). The finite volume method consists of integrating the equations of the model (contrary to the finite element method which is based on integrating the equations of the weak formulation). Applying the finite volume method, we integrate the equations of system (2.29) over the set $K \times [t_n, t_{n+1}]$ with $K \in \mathcal{T}$ and $n \in \{0, \cdots, N+1\}$, further by the use of the Green-Gauss formula for the divergence operators, the finite volume consists of approximating the resulting normal fluxes across the interfaces $\{\sigma_{KL}\}_{L \in \mathcal{N}(K)}$ of the control volume K.

Time evolution terms

For the time evolution terms of system (2.29), and using the Taylor expansion, one gets

$$\begin{aligned}\int_{t_n}^{t_{n+1}} \int_K \partial_t w\,(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t &\approx \Delta t \int_K \partial_t w\,(\mathbf{x}, t_n)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &= \int_K \left(w\,(\mathbf{x}, t_{n+1}) - w\,(\mathbf{x}, t_n)\right)\mathrm{d}\mathbf{x} = |K|\left(w_K^{n+1} - w_K^n\right),\end{aligned}$$

with $w \equiv u$ or v.

Diffusion terms

We consider the diffusive term of the first equation of system (2.29), one has

$$\int_{t_n}^{t_{n+1}} \int_K \mathrm{div}\,\left(\nabla A\,(u)\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t = \sum_{L \in \mathcal{N}(K)} \int_{t_n}^{t_{n+1}} \int_{\sigma_{KL}} \nabla A\,(u) \cdot \eta_{KL}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t,$$

where, $\mathrm{d}\sigma(\mathbf{x})$ is the Lebesgue measure on the edge σ_{KL}.
Using again Taylor's formula as well as the orthogonality condition (i.e. $\mathbf{x}_L - \mathbf{x}_K = d_{KL}\eta_{KL}$), one can deduce that the flux $\nabla A\,(u) \cdot \eta_{KL}$ can be approximated on the interface σ_{KL} by

$$\nabla A\,(u) \cdot \eta_{KL} \approx \frac{u\,(\mathbf{x}_L) - u\,(\mathbf{x}_K)}{d_{KL}}. \tag{2.39}$$

Using the flux approximation (2.39), one can deduce the following approximation of the diffusion term

$$\int_{t_n}^{t_{n+1}} \int_K \mathrm{div}\,\left(\nabla A\,(u)\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \approx \Delta t \sum_{L \in \mathcal{N}(K)} \tau_{KL}(A\,(u_L) - A\,(u_K)),$$

where, $\tau_{KL} := \dfrac{|\sigma_{KL}|}{d_{KL}}$ represents the transmissibility through the interface σ_{KL}. In the same manner, we treat the diffusion term of the second equation of system (2.29) to get

$$\int_{t_n}^{t_{n+1}} \int_K \mathrm{div}\,\left(\nabla v\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \approx \Delta t \sum_{L \in \mathcal{N}(K)} \tau_{KL}(v_L - v_K).$$

Convection terms

Let us consider the convective term of the first equation of system (2.29). For the simplicity, we denote by Ψ the function defined by $\Psi(u) := u\,q(u)\,\chi$. We note that the discretization of the convective term is slightly different from that one of the diffusion term, here we have to approximate $\Psi(u)\,\nabla v \cdot \eta_{KL}$ at the interface σ_{KL} and at time t_{n+1}. The classical choice is to approach the flux using an upwind finite volume scheme. This technique consists of approximating the flux by means of the values u_K, u_L, and $\mathrm{d}V_{KL} := \tau_{KL}(v_L - v_K)$ which are available in the neighborhood of the interface σ_{KL}. To do this, we use a numerical convection flux function G of arguments $(a,b,c) \in \mathbb{R}^3$ which is required to satisfy the properties :

$$\begin{cases} \text{(a) } G(\cdot,b,c) \text{ is nondecreasing for all } b,c \in \mathbb{R}, \\ \text{and } G(a,\cdot,c) \text{ is nonincreasing for all } a,c \in \mathbb{R}; \\ \text{(b) } G(a,b,c) = -G(b,a,-c) \text{ for all } a,b,c \in \mathbb{R}; \\ \text{(c) } G(a,a,c) = \Psi(a)\,c \text{ for all } a,c \in \mathbb{R}; \\ \text{(d) there exists } C>0 \text{ such that} \\ \quad \forall\, a,b,c \in \mathbb{R} \quad |G(a,b,c)| \le C\,(|a|+|b|)\,|c|; \\ \text{(e) there exists a modulus of continuity } \omega : \mathbb{R}^+ \to \mathbb{R}^+ \text{ such that} \\ \quad \forall\, a,b,a',b',c \in \mathbb{R} \quad |G(a,b,c) - G(a',b',c)| \le |c|\,\omega(|a-a'|+|b-b'|). \end{cases} \tag{2.40}$$

Remark 2. Note that the assumptions on the dependence of G on a, b are standard (see, e.g.[31]). For instance, the assumptions (a), (b), and (c) ensure respectively the maximum principle on the discrete solutions, the conservativity of the numerical flux, and the consistency of the numerical flux on each interface. Practical examples of numerical convective flux functions can be found in [31].
In our context, one possibility to construct the numerical flux G satisfying (2.40) is to split Ψ into the nondecreasing part $\Psi_\uparrow$ and the nonincreasing part $\Psi_\downarrow$:

$$\Psi_\uparrow(z) := \int_0^z \left(\Psi'(s)\right)^+ \mathrm{d}s, \qquad \Psi_\downarrow(z) := -\int_0^z \left(\Psi'(s)\right)^- \mathrm{d}s.$$

Herein, $s^+ = \max(s,0)$ and $s^- = \max(-s,0)$. Then we take

$$G(a,b;c) = c^+\Big(\Psi_\uparrow(a) + \Psi_\downarrow(b)\Big) - c^-\Big(\Psi_\uparrow(b) + \Psi_\downarrow(a)\Big). \tag{2.41}$$

Discretization of system (2.29)

We are now in a position to discretize problem (2.29). We denote by $\mathcal{D}$ an admissible discretization of Q_{t_f}, which consists of an admissible mesh of Ω and a uniform time step $\Delta t > 0$. We give to the parameter h the sense of

$$\max\Big\{\Delta t\,,\ \max_{K\in\mathcal{T}} \mathrm{diam}(K)\,,\ \max_{K\in\mathcal{T}}\max_{L\in\mathcal{N}(K)} d_{KL}\Big\}.$$

A finite volume scheme for the discretization of the problem (2.29) is given by the following set of equations : for all $K \in \mathcal{T}$

$$u_K^0 = \frac{1}{|K|}\int_K u_0(\mathbf{x})\,\mathrm{d}\mathbf{x}, \qquad v_K^0 = \frac{1}{|K|}\int_K v_0(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{2.42}$$

and for all $K \in \mathcal{T}$ and n$\in [\![0 \cdots N]\!]$

$$
\begin{aligned}
&|K| \frac{u_K^{n+1} - u_K^n}{\Delta t} - d_1 \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{d_{KL}} \left(A\left(u_L^{n+1}\right) - A\left(u_K^{n+1}\right)\right) \\
&\qquad + \sum_{L \in N(K)} G\left(u_K^{n+1}, u_L^{n+1}; \mathrm{d}V_{KL}^{n+1}\right) = |K| \mu u_K^{n+1} \left(1 - \frac{u_K^{n+1}}{u_c}\right), \qquad (2.43) \\
&|K| \frac{v_K^{n+1} - v_K^n}{\Delta t} - d_2 \sum_{L \in N(K)} \frac{|\sigma_{KL}|}{d_{KL}} \left(v_L^{n+1} - v_K^{n+1}\right) = |K| \left(\alpha u_K^n - \beta v_K^{n+1}\right),
\end{aligned}
$$

with the discrete unknowns $U = \left(u_K^{n+1}\right)_{K \in \mathcal{T}}$ and $V = \left(v_K^{n+1}\right)_{K \in \mathcal{T}}, n \in [\![0 \cdots N]\!]$.

Notice that the discrete zero-flux boundary conditions are implicitly contained in equations (2.43). Indeed, we have for all $K \in \mathcal{T}$, the contribution of $\partial\Omega \cap \partial T$ to the approximation of $\displaystyle\int_{\partial K} \nabla v \cdot \eta$, $\displaystyle\int_{\partial K} \nabla A(u) \cdot \eta$ is zero, in compliance with the last equation of system (2.29).
The convergence analysis of the finite volume scheme (2.42)–(2.43) is carried out in [36] for the nondegenerate case and in [7] for the degenerate case. We refer to those references to make sure that the scheme (2.42)–(2.43) converges towards the weak solution of problem (2.29).

2.5.3 Numerical results

In this section, we show two numerical simulations of model (2.29) in a two dimensional space to illustrate our theoretical results and demonstrate the patterns generated by the system. To do this, we use the Newton algorithm to approach the solution of the nonlinear system defined by the first equation of (2.43). This algorithm is coupled with a bigradient method to solve linear systems arising from the Newton algorithm process.
We simulate the model in a two dimensional domain $\Omega = (0, 5) \times (0, 5)$ for which we consider a nonuniform admissible grid (see Figure 2.6). We consider the following data for the numerical

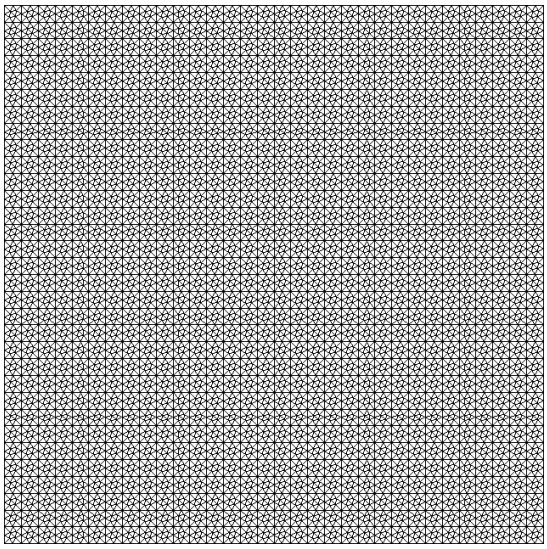


FIGURE 2.6 – Admissible mesh with 14 346 triangles.

test $d_1 = 0.1$, $d_2 = 1.0$, $u_c = 0.25$, $\bar{u} = 1.0$, $\mu = 0.5$, $\alpha = 10.0$, $\beta = 10.0$, $\chi = 20$, and $\gamma = 2$.

Further, we consider a small time step $\Delta t = 0.05$ and in the definition of the numerical fluxes G defined by (2.40), we consider

$$\Psi_{\uparrow}(z) = \Psi(\min\{z, u_m\}) \text{ and } \Psi_{\downarrow}(z) = \Psi(\max\{z, u_m\}) - \Psi(u_m)$$

where $u_m = \dfrac{\overline{u}}{\sqrt[\gamma]{\gamma+1}}$.

Initial conditions and boundary conditions

Initially the cell density is setting as a small perturbation around the homogeneous steady state ; in our test, we take $u(0, \mathbf{x}, \mathbf{y}) = u_s + \varepsilon \cos(n\pi\mathbf{x}) \cos(m\pi\mathbf{x})$ where $u_s = u_c = 0.25$, $\varepsilon = 0.001$, $m = 6$, and $n = 0$. While for the chemoattractant, we consider a random perturbation around the steady state. Fig. 2.7 shows the plot of the initial condition for the cell density and for the chemoattractant ; the red regions correspond to the highest cell density so the cells are aggregated into four columns "*four stripes*". For the boundary conditions, we impose zeros-flux boundary conditions.

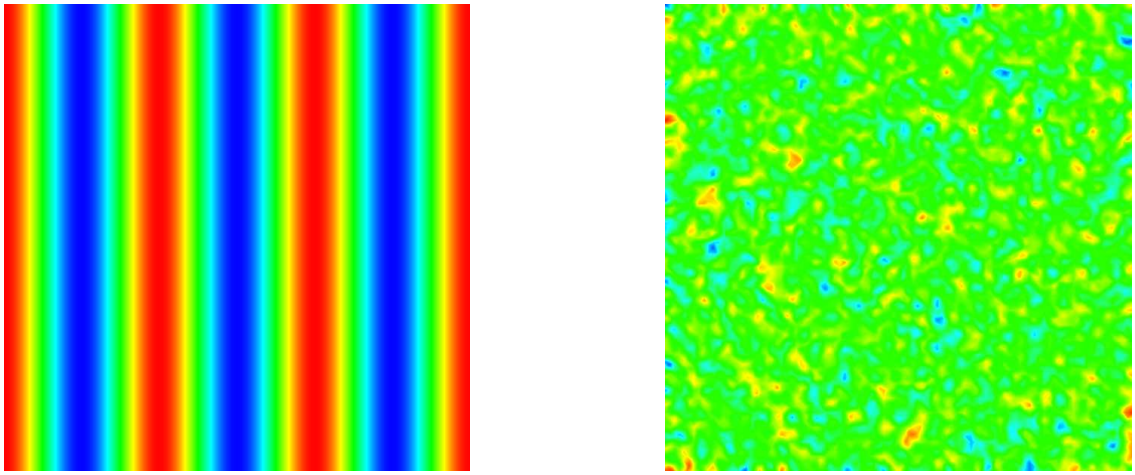


FIGURE 2.7 – Initial condition for each of the cell density $0.249 \leq u \leq 0.251$ (*left*) with a small perturbation around the steady state, and for the chemoattractant (*right*) with random perturbation around v_s.

Spatial evolution of the cell density

Figures 2.8-2.10 show the evolution of the cell density in space for different moments. We note that at $t = 0.1\text{s}$ the cell density u begins to disperse with response to the gradient of the chemoattractant v but the cell density remains distributed around the four columns. Then after half a second, the spatial patterns begin to form and the aggregations start to merge and emerge ($t = 1\text{s}$) and then to form uniform spot patterns as at $t = 14\text{s}$. Next, after 228s, these spot patterns stabilize and finally form 13 spatially inhomogeneous stable patterns as we can see in Fig 2.10, where the two plots show the same patterns for different moments.

Time evolution of the cell density

Here we are interested in the time evolution of the cell density at fixes points in the last plot of Fig. 2.10. Indeed, we want to show that the cell density stabilize at a certain moment ; thus we verify the theoretical results about the pattern formation using an appropriate numerical scheme.

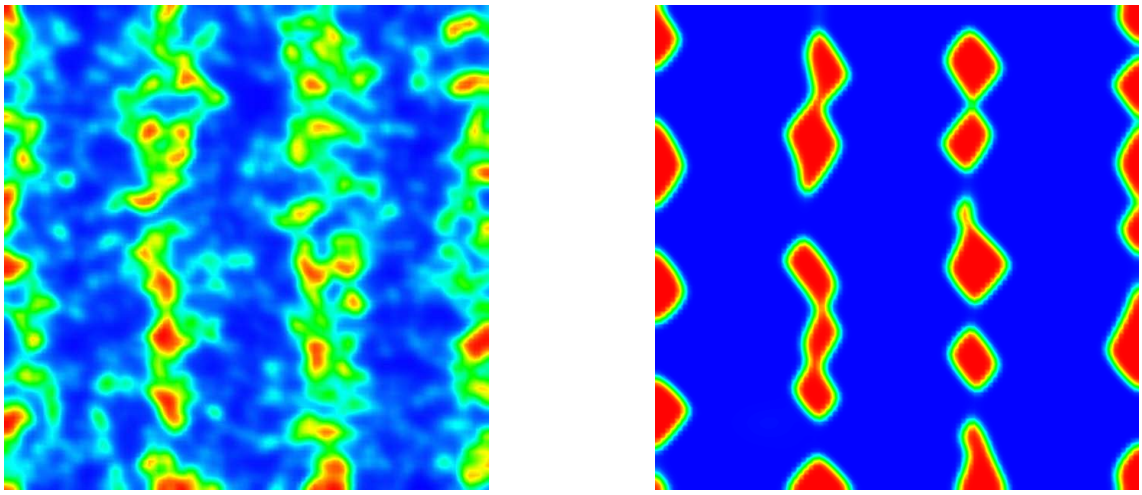


FIGURE 2.8 – Pattern formation for the full chemotaxis model (2.29) at time $t = 0.1\,\mathrm{s}$ with $1.76 \times 10^{-2} \leq u \leq 0.97$ (*left*) and at time $t = 0.6\mathrm{s}$ with $3.65 \times 10^{-5} \leq u \leq 0.99$ (*right*).

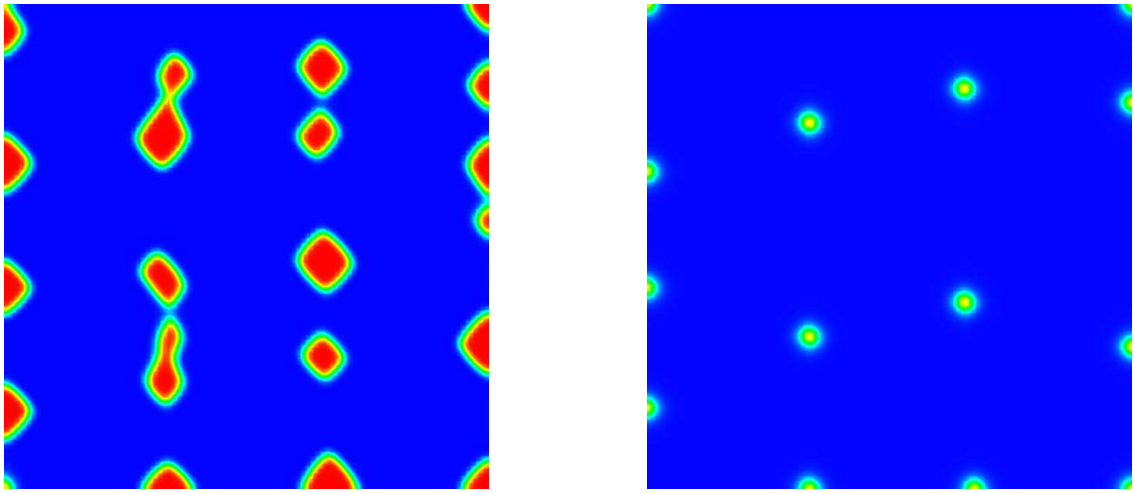


FIGURE 2.9 – Pattern formation for the full chemotaxis model (2.29) at time $t = 1\mathrm{s}$ with $2.46 \times 10^{-6} \leq u \leq 0.99$ (*left*) and at time $t = 14\mathrm{s}$ with 1.3×10^{-3} (*right*).

Figure 2.11 shows the evolution of the cell density at point $P_1\,(3.375; 2.025)$ in the blue line and at point $P_2\,(4:975; 2:825)$ in the red line. We observe that the cell density at the two points increases and decreases with response to the gradient of the chemoattractant which plays an essential role to stop the aggregation of the cells.
Next, the cell density stabilizes for both points when t is greater or equal to $178\mathrm{s}$, we get the same result for the others spot patterns ; we do not provide them here for convenience. Consequently, this stabilizing proves that the volume-filling chemotaxis model (2.29) generates stationary spatial patterns.

Test2 Now we consider a second numerical test for the generation of spatial patterns of the volume-filling chemotaxis model. Here, for the initial condition of the cell density, we take a random distribution of the form $u\,(0, \mathbf{x}, \mathbf{y}) = u_s + \varepsilon \times \mathrm{rand}(0, 1)$ where $u_s = 0.25$, $\varepsilon = 0.01$, and rand is function which returns a pseudo-random number from a uniform distribution between 0 and 1. We take the same data as the previous test except for the domain size which replaced by $\Omega = (0, 10) \times (0, 10)$. Figures 2.12–2.14 show the evolution of the cell density in space for

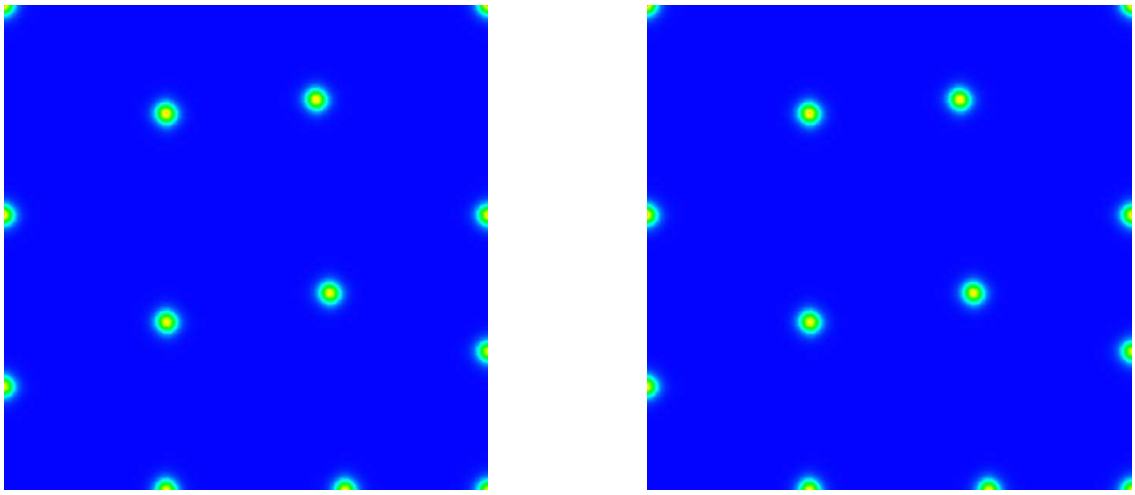


FIGURE 2.10 – Pattern formation for the full chemotaxis model (2.29) at time $t = 242\,\mathrm{s}$ with $1.45 \times 10^{-3} \leq u \leq 0.79$ (*left*) and at time $t = 316\,\mathrm{s}$ with $1.45 \times 10^{-3} \leq u \leq 0.79$ (*right*).

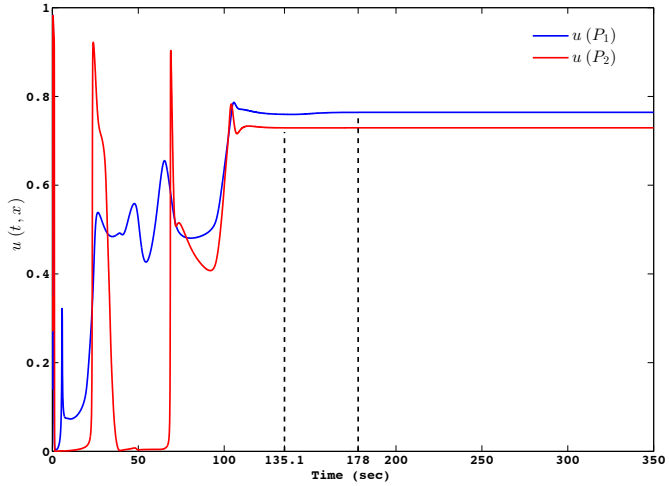


FIGURE 2.11 – Evolution of the cell density with respect to the time at two different points P_1 and P_2.

different moments. As before, we observe the emerging process and then we get stationary spot patterns.

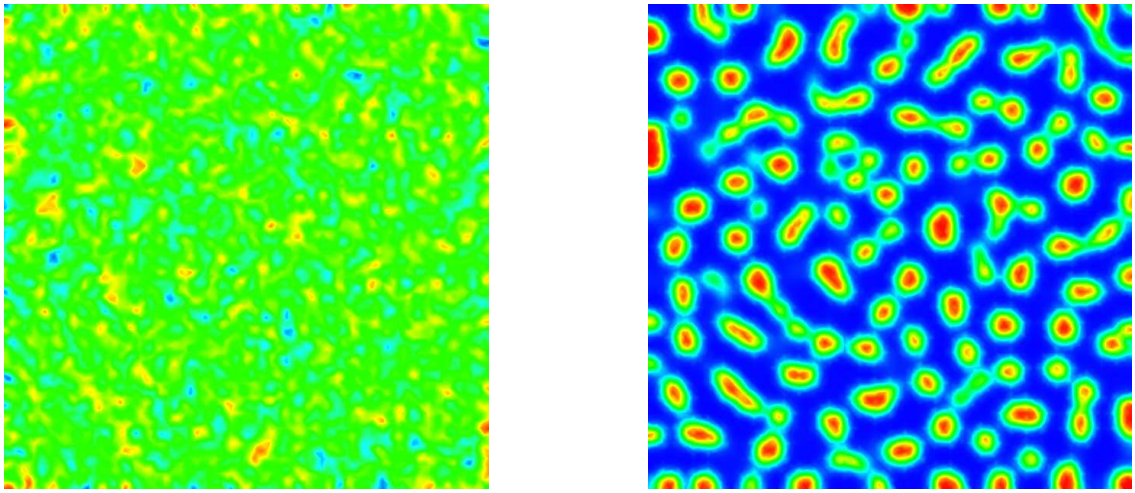


FIGURE 2.12 – Pattern formation for the full chemotaxis model (2.29) at time $t = 0$s with $0.25 \leq u \leq 0.2599$ (*left*) and at time $t = 4$s with $3.3 \times 10^{-3} \leq u \leq 0.99$ (*right*).

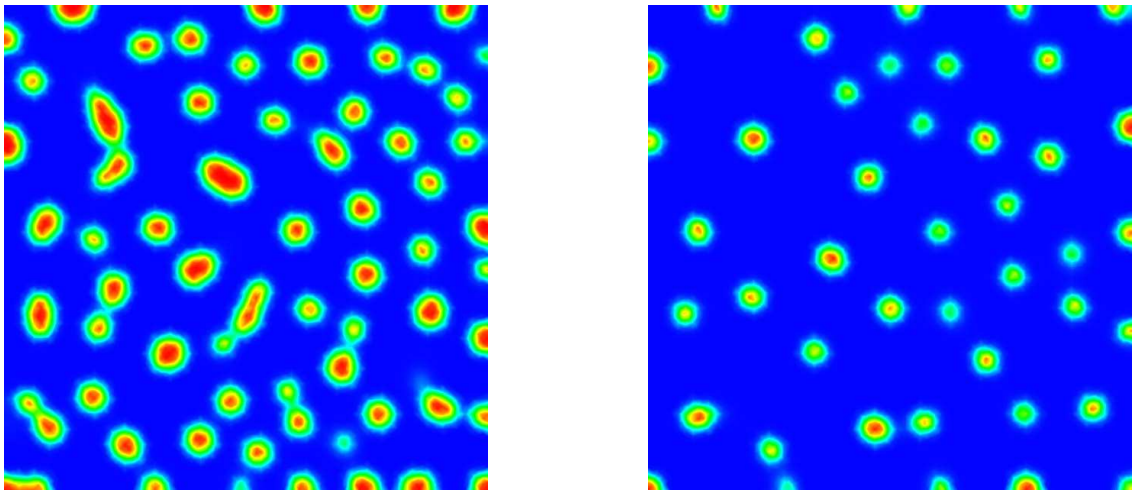


FIGURE 2.13 – Pattern formation for the full chemotaxis model (2.29) at time $t = 10$ s with $1.87 \times 10^{-5} \leq u \leq 0.99$ (*left*) and at time $t = 240$s with $2.14 \times 10^{-5} \leq u \leq 0.97$ (*right*).

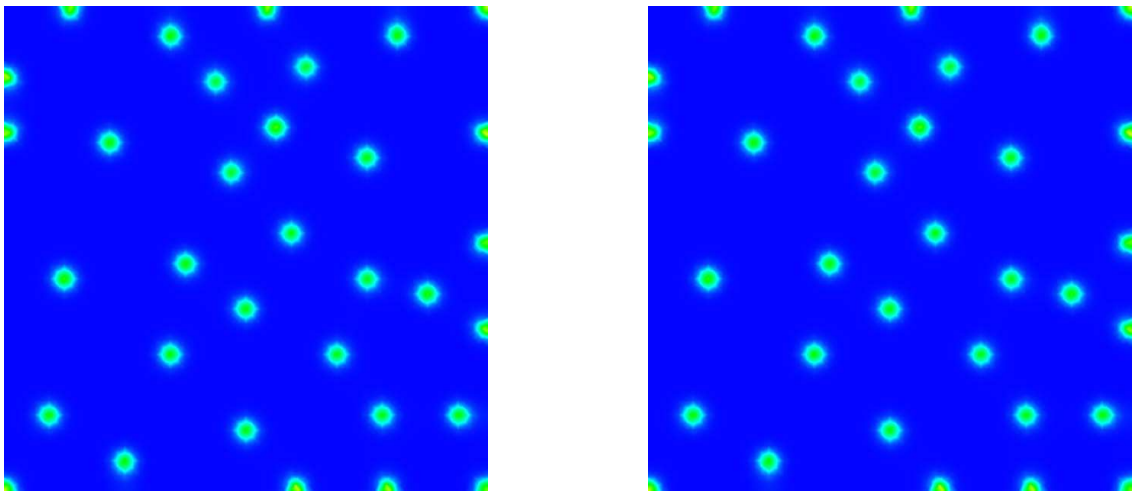


FIGURE 2.14 – Pattern formation for the full chemotaxis model (2.29) at time $t = 400$ s with $1.07 \times 10^{-3} \leq u \leq 0.691$ (*left*) and at time $t = 500$s with $1.07 \times 10^{-3} \leq u \leq 0.691$ (*right*).

3

CVFE scheme for the capture of patterns for a volume-filling chemotaxis model

Sommaire

3.1 Introduction

Patterns are the solutions of a reaction–diffusion system which are stable in time and stationary inhomogeneous in space, while pattern formation in mathematics refers to the process that, by changing a bifurcation parameter, the spatially homogeneous steady states lose stability to spatially inhomogeneous perturbations, and stable inhomogeneous solutions arise.

The pattern formation has been successfully applied to bacteria (see e.g. [77]) where we investigate specific and necessary parameters to obtain stationary distribution of the disease. Also, it

has been applied to skin pigmentation patterns [59] to understand the diversity of patterns on the animal coat pattern, and many other examples.

The pattern formation depends on two key properties : the first is to apply the seminal idea of Turing [76] for a reaction–diffusion system and consequently determine the bifurcation parameters for the generation of stationary inhomogeneous spatial patterns (also called Turing Patterns), and the second is to apply a robust scheme to numerically investigate and capture the generation of spatio-temporal patterns. One of the most popular reaction–diffusion systems that can generate spatial patterns is the chemotaxis model.

Chemotaxis is the feature movement of a cell along a chemical concentration gradient either towards the chemical stimulus, and in this case the chemical is called chemoattractant, or away from the chemical stimulus and then the chemical is called chemorepellent. The mathematical analysis of chemotaxis models shows a plenitude of spatial patterns such as the chemotaxis models applied to skin pigmentation patterns [63, 67] that lead to aggregations of one type of pigment cell into a striped spatial pattern. Other models have been successfully applied to the aggregation patterns in an epidemic disease [9], tumor growth [19], angiogenesis in tumor progression [14], and many other examples. Theoretical and mathematical modeling of chemotaxis dates to the pioneering works of Patlak in the 1950s [69] and Keller and Segel in the 1970s [51, 52]. The review article by Horstmann [47] provides a detailed introduction into the mathematics of the Keller-Segel model for chemotaxis.

In this chapter, we present and study a numerical scheme for the capture of spatial patterns for a nonlinear degenerate volume-filling chemotaxis model over a general mesh, and with inhomogeneous and anisotropic diffusion tensors. Recently, the convergence analysis of a finite volume scheme for a degenerate chemotaxis model over a homogeneous domain has been studied by Andreianov *et al.* [7], where the diffusion tensor is considered to be proportional to the identity matrix, and the mesh used for the space discretization is assumed to be admissible in the sense of satisfying the orthogonality condition as in [31]. The upwind finite volume method used for the discretization of the convective term ensures stability and is extremely robust and computationally inexpensive. However, standard finite volume scheme does not permit handling anisotropic diffusion on general meshes, even if the orthogonality condition is satisfied. The reason for this is that there is no straightforward way to apply the finite volume scheme to problems with full diffusion tensors. Various "multi-point" schemes, where the approximation of the flux through an edge involves several scalar unknowns, have been proposed, see for e.g. [33, 32] for the so-called SUCHI scheme, [25, 22] for the so-called gradient scheme, and [3] for the development of the so-called DDFV schemes. However, such schemes require using more points than the classical 4 points for triangular meshes in space dimension two, making the schemes less robust and more susceptible to numerical instabilities.

To handle the discretization of the anisotropic diffusion, it is well-known that the finite element method allows for an easy discretization of the diffusion term with a full tensor. However, it is well-established that numerical instabilities may arise in the convection-dominated case. To avoid these instabilities, the theoretical analysis of the *control volume finite element method* has been carried out for the case of degenerate parabolic problems with full diffusion tensors. Schemes with mixed conforming piecewise linear finite elements on triangles for the diffusion term and finite volumes on dual elements were proposed and studied in [15, 20, 29, 2, 17] for fluid mechanics equations, are indeed quite efficient.

Afif and Amaziane analyzed in [2] the convergence of a vertex-centered finite volume scheme for a nonlinear and degenerate convection-diffusion equation modeling a flow in porous media and without reaction term. This scheme consists of a discretization of the Laplacian by the piecewise

linear conforming finite element method (see also [64, 37]), the effectiveness of this scheme was tested in benchmarks of FVCA series of conferences [42]. Cariaga *et al.* in [17] considered the same scheme for a reaction–diffusion-convection system, where the velocity of the fluid flow is considered to be constant in the convective term.

The intention of this chapter is to extend the ideas of [7, 2, 17] to a fully nonlinear degenerate parabolic system modeling the effect of volume-filling for chemotaxis. In order to discretize this class of systems, we discretize the diffusion term by means of piecewise linear conforming finite element. The other terms are discretized by means of a finite volume scheme on a dual mesh (Donald mesh), with an upwind discretization of the numerical flux of the convective term to ensure the stability and the maximum principle of the scheme, where the dual mesh is constructed around each vertex of every triangle of the primary mesh.

The rest of this chapter is organized as follows. In Section 3.2, we introduce the chemotaxis model based on realistic biological assumptions, which incorporates the effect of volume-filling mechanism and leads to a nonlinear degenerate parabolic system. In Section 3.3, we derive the *control volume finite element scheme*, where an upwind finite volume scheme is used for the approximation of the convective term, and a standard P1-finite element method is used for the diffusive term. In Section 3.4, by assuming that the transmissibility coefficients are nonnegative, we prove the discrete maximum principle and give the *a priori* estimates on the discrete solutions. In Section 3.5, we show the compactness of the set of discrete solutions by deriving estimates on difference of time and space translates for the approximate solutions. Next, in Section 3.6, using the Kolmogorov relative compactness theorem, we prove the convergence of a sequence of the approximate solutions, and we identify the limits of the discrete solutions as weak solutions of the parabolic system proposed in Section 3.2. In the last section, we present some numerical simulations to capture the generation of spatial patterns for the volume-filling chemotaxis model with different tensors. These numerical simulations are obtained with our *control volume finite element scheme*.

3.2 Volume-filling chemotaxis model

We are interested in the *control volume finite element scheme* for a nonlinear, degenerate parabolic system formed by reaction–diffusion–convection equations. This system is complemented with homogeneous zero flux boundary conditions, which correspond to the physical behavior of the cells and the chemoattractant. The modified Keller-Segel system that we consider here, is very similar to that of Andreianov *et al.* [7], to which we have added tensors for the diffusion terms. Specifically, we consider the following system :

$$\begin{cases} \partial_t u - \operatorname{div}\left(\Lambda(\mathbf{x})\, a(u)\, \nabla u - \Lambda(\mathbf{x})\, \chi(u)\, \nabla v\right) = f(u) & \text{in } Q_{t_f}, \\ \partial_t v - \operatorname{div}\left(D(\mathbf{x})\, \nabla v\right) = g(u, v) & \text{in } Q_{t_f}, \end{cases} \tag{3.1}$$

with the boundary conditions on $\Sigma_{t_f} := \partial\Omega \times (0, t_f)$ given by

$$\left(\Lambda(\mathbf{x})\, a(u)\, \nabla u - \Lambda(\mathbf{x})\, \chi(u)\, \nabla v\right) \cdot \mathbf{n} = 0, \qquad D(\mathbf{x})\, \nabla v \cdot \mathbf{n} = 0, \tag{3.2}$$

where, $\mathbf{n}$ is the unit outward normal vector at the boundary $\partial\Omega$ of the domain Ω.
The initial conditions are given by :

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad v(\mathbf{x}, 0) = v_0(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{3.3}$$

Herein, $Q_{t_f} := \Omega \times (0, t_f)$, $t_f > 0$ is a fixed time, and Ω is an open bounded polygonal domain in $\mathbb{R}^2$, with Lipschitz boundary $\partial\Omega$. The initial conditions u_0 and v_0 satisfy : $u_0, v_0 \in$

$L^{\infty}(\Omega)$ such that $0 \leq u_0(\mathbf{x}) \leq 1$ and $v_0(\mathbf{x}) \geq 0$, for all $\mathbf{x} \in \Omega$.
In the above model, the density of the cell-population and the chemoattractant (or repellent) concentration are represented by $u = u(\mathbf{x}, t)$ and $v = v(\mathbf{x}, t)$ respectively. Next, $a(u)$ is a density-dependent diffusion coefficient, and $\Lambda(\mathbf{x})$ is the diffusion tensor in a heterogeneous medium. Furthermore, the function $\chi(u)$ is the chemoattractant sensitivity, and $D(\mathbf{x})$ is the diffusion tensor for v. The function $f(u)$ describes cell proliferation and cell death, it is usually considered to follow the logistic growth with certain carrying capacity u_c which represents the maximum density that the environment can support (e.g. see [10]). The function $g(u, v)$ describes the rates of production and degradation of the chemoattractant ; here, we assume it is the linear function given by

$$g(u, v) = \alpha u - \beta v, \qquad \alpha, \beta \geq 0. \tag{3.4}$$

Painter and Hillen [66] introduced the mechanistic description of the volume-filling effect. In the volume-filling effect, it is assumed that particles have a finite volume and that cells cannot move into regions that are already filled by other cells. First, we give a brief derivation of the model below, where in addition we consider the elastic cell property ; that is, we consider that the cells are deformable and elastic and can squeeze into openings.

The derivation of the model begins with a master equation for a continuous-time and discrete space-random walk introduced by Othmer and Stevens [74], that is

$$\frac{\partial u_i}{\partial t} = \mathcal{C}_{i-1}^{+} u_{i-1} + \mathcal{C}_{i+1}^{-} u_{i+1} - \left(\mathcal{C}_i^{+} + \mathcal{C}_i^{-}\right) u_i, \tag{3.5}$$

where $\mathcal{C}_i^{\pm}$ are the transitional probabilities per unit of time for one-step jump to $i \pm 1$. Herein, we shall equate the probability distribution above with the cell density.

In the volume-filling approach, and in the context of chemotaxis, the probability of a cell making a jump is assumed to depend on additional factors, such as the external concentration of the chemotactic agent and the availability of space into which the cells can squeeze and move. Therefore, we consider in the transition probability the fact that the cells can detect a local gradient as well as squeeze into openings. We take

$$\mathcal{C}_i^{\pm} = q\left(u_{i\pm1}\right)\left(\theta + \delta\left[\tau\left(v_{i\pm1}\right) - \tau\left(v_i\right)\right]\right), \tag{3.6}$$

where $q(u)$ is a nonlinear function representing the squeezing probability of a cell finding space at its neighboring location, θ and δ are constants, and τ represents the mechanism of the signal detection of the chemical concentration (for more details see [66, 74, 45]). It is assumed that only a finite number of cells, $\overline{u}$, can be accommodated at any site, and the function q is stipulated by the following condition :

$$q(\overline{u}) = 0 \quad \text{and} \quad 0 < q(u) \leq 1 \quad \text{for } 0 \leq u < \overline{u}.$$

Clearly, a possible choice for the squeezing probability q is a nonlinear function (see [78] for more details), defined by

$$q(u) = \begin{cases} 1 - \left(\dfrac{u}{\overline{u}}\right)^{\gamma}, & 0 \leq u \leq \overline{u}, \\ 0, & u > \overline{u}, \end{cases} \tag{3.7}$$

where $\gamma \geq 1$ denotes the squeezing exponent. The case $\gamma = 1$ corresponds to the interpretation that cells are solid blocks. However, the cells are elastic and can squeeze into openings. Thus the squeezing probability should be considered as a nonlinear function.

Substituting equation (3.6) into the master equation (3.5) and assuming that the cell density can diffuse in a heterogeneous manner in the space we get the first equation of system (3.1), with the associated coefficients

$$a(u) = d_1\left(q(u) - q'(u)\, u\right), \quad \chi(u) = \zeta u q(u), \tag{3.8}$$

where d_1 and ζ are two positive constants.

In this chapter, we are interested in system (3.1) modeling the volume-filling chemotaxis process in the general case and for which we set $\overline{u} = 1$. Furthermore, we assume that the functions f and χ are continuous and satisfy :

$$f(0) = f(\overline{u}) = 0, \qquad \chi(0) = \chi(\overline{u}) = 0. \tag{3.9}$$

Now, we state the main assumptions made about the system (3.1)–(3.3)

(A1) $a : [0,1] \longrightarrow \mathbb{R}$ is a continuous function such that : $a(0) \geq 0$, $a(1) \geq 0$ and $a(u) > 0$ for all $0 < u < 1$.

(A2) $\chi : [0,1] \longrightarrow \mathbb{R}$ is a differentiable function such that : $\chi(0) = \chi(1) = 0$ and $\chi(u) > 0$ for all $0 < u < 1$.

(A3) The diffusion tensors Λ and D are two bounded, uniformly positive symmetric tensors on Ω, that is : $\forall \mathbf{w} \neq 0$, there exists two nonnegative constants T_- and T_+ such that $0 < T_- |\mathbf{w}|^2 \leq \langle T(\mathbf{x})\mathbf{w}, \mathbf{w}\rangle \leq T_+ |\mathbf{w}|^2 < \infty, \quad T = \Lambda$ or D.

(A4) The cell proliferation function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous function such that : $f(0) \geq 0$ and $f(1) \leq 0$.

(A5) The initial conditions u_0 and v_0 are two functions lying into $L^\infty(\Omega)$ such that : $0 \leq u_0 \leq 1$ and $v_0 \geq 0$.

In the sequel, we use the nonlinear Lipschitz continuous nondecreasing function $A : [0,1] \longrightarrow \mathbb{R}$ defined by

$$A(u) = \int_0^u a(s)\, \mathrm{d}s. \tag{3.10}$$

Definition 3.1 (weak solution)**.** Under the assumptions (A1)–(A5). The couple of measurable functions (u, v) is said to be a weak solution of the system (3.1)–(3.3) if

$$\begin{aligned}
&0 \leq u(\mathbf{x}, t) \leq 1, \ v(\mathbf{x}, t) \geq 0 \text{ for almost everywhere } (\mathbf{x}, t) \in Q_{t_\mathrm{f}}, \\
&A(u) \in L^2\left(0, t_\mathrm{f}; H^1(\Omega)\right), \\
&v \in L^\infty(Q_{t_\mathrm{f}}) \cap L^2\left(0, t_\mathrm{f}; H^1(\Omega)\right),
\end{aligned}$$

and for all $\varphi, \psi \in \mathcal{D}\left(\overline{\Omega} \times [0, t_\mathrm{f})\right)$

$$\begin{aligned}
-\int_\Omega u_0(\mathbf{x})\, \varphi(\mathbf{x}, 0)\, \mathrm{d}\mathbf{x} - \iint_{Q_{t_\mathrm{f}}} u \partial_t \varphi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t + \iint_{Q_{t_\mathrm{f}}} \Lambda(\mathbf{x}) \nabla A(u) \cdot \nabla \varphi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\
- \iint_{Q_{t_\mathrm{f}}} \Lambda(\mathbf{x})\, \chi(u) \nabla v \cdot \nabla \varphi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t = \iint_{Q_{t_\mathrm{f}}} f(u)\, \varphi(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t, \\
-\int_\Omega v_0(\mathbf{x})\, \psi(\mathbf{x}, 0)\, \mathrm{d}\mathbf{x} - \iint_{Q_{t_\mathrm{f}}} v \partial_t \psi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\
+ \iint_{Q_{t_\mathrm{f}}} D(\mathbf{x}) \nabla v \cdot \nabla \psi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t = \iint_{Q_{t_\mathrm{f}}} g(u, v)\, \psi(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t.
\end{aligned}$$

3.3 CVFE discretization of the continuous problem

We are placed in the case where the boundary of the domain occupied by the cells and the chemeoattractant is fixed over the time. We consider that $\Omega \subset \mathbb{R}^2$ is the domain occupied by the cells, and that $\Omega_h \subset \Omega$ is the approached polygonal domain of Ω. The construction of the approached solution requires the introduction of two different space discretizations of the domain Ω_h. The first discretization namely the *primal mesh* consists of simple polygons that form a partition of the domain Ω_h. The approach adopted here is oriented to the summits ("Vertex Centered approach"), the objective of this technique is to construct the solution at the summits of the *primal mesh*. To do this, a new partition of the domain is defined in such a way that every vertex of the *primal mesh* is included into only a polygon of the new partition. The mesh resulting from the second partition is called *dual mesh* and the polygons that compose it are called *dual control volumes.*

Definition 3.2. (*Primal and dual mesh*). Let Ω be an open bounded polygonal connected subset of $\mathbb{R}^2$. A primal finite volume mesh of Ω is a triplet $(\mathcal{T}, \mathcal{E}, \mathcal{P})$, where $\mathcal{T}$ is a family of disjoint open polygonal convex subsets of Ω called control volumes, $\mathcal{E}$ is a family of subsets of $\overline{\Omega}$ contained in straights of $\mathbb{R}^2$ with strictly positive one-dimensional measure, called the edges of the control volumes, and $\mathcal{P}$ is a family of points of Ω satisfying the following properties :

1. The closure of the union of all the control volumes is $\overline{\Omega}$, i.e. $\overline{\Omega} = \bigcup_{K\in\mathcal{T}} \overline{K}$.

2. For any $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \overline{K}\backslash K = \bigcup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$. Furthermore, $\mathcal{E} = \bigcup_{K\in\mathcal{T}} \mathcal{E}_K$.

3. There exists a Donald dual mesh $\mathcal{M} := \{M_i, i = 1, \ldots, N_s\}$ associated with the triangulation $\mathcal{T} := \{K_i, i = 1, \ldots, N_e\}$. For each triangle $K \in \mathcal{T}$, we connect the barycenter $\mathbf{x}_K$ with the midpoint of each edge $\sigma \in \mathcal{E}_K$, and thus the barycenter of $K \in \mathcal{T}$ is such that $\mathbf{x}_K := \bigcap_{M\cap K\neq\emptyset} \partial M \in K$. We denote by $\mathbf{x}_M$ the center of each dual control volume $M \in \mathcal{M}$ defined by $\mathbf{x}_M := \bigcap_{K\cap M\neq\emptyset} \partial K \in M$. For each interface of the dual control volume M, we denote by $\sigma^K_{M,M'} := \partial M \cap \partial M' \cap K$ the line segment between the points $\mathbf{x}_K$ and the midpoint of the line segment $[\mathbf{x}_M, \mathbf{x}_{M'}]$ and let $\pounds := \{\sigma \in \partial M\backslash\partial\Omega, M \in \mathcal{M}\}$ be the set of all interior sides. Finally we denote by $\mathcal{M}^{\text{int}}$ and $\mathcal{M}^{\text{ext}}$ the set of all interior and all boundary dual control volumes respectively. We refer to Fig. 3.1 for an illustration of the primal triangular mesh $\mathcal{T}$ and its corresponding Donald dual mesh $\mathcal{M}$.

In the sequel, we use the following notations. For any $M \in \mathcal{M}$, $|M|$ is the area of M. The set of neighbors of M is denoted by $\mathcal{N}(M) := \left\{M' \in \mathcal{M} / \exists\sigma \in \pounds, \overline{\sigma} = \overline{M} \cap \overline{M'}\right\}$, and we designate by $d_{M,M'}$ the distance between the centers of M and M'. We define the mesh size by $h := \text{size}(\mathcal{M}) = \sup_{M\in\mathcal{M}} \text{diam}(M)$ and make the following shape regularity assumption on the family of triangulations $\{\mathcal{T}_h\}_h$:

$$\text{There exists a positive constant } \kappa_{\mathcal{T}} \text{ such that : } \quad \min_{K\in\mathcal{T}_h} \frac{|K|}{\text{diam}\,(K)^2} \geq \kappa_{\mathcal{T}}, \qquad \forall h > 0. \quad (3.11)$$

For the time discretization, we do not impose any restriction on the time step, for that we consider a uniform time step $\Delta t \in (0, t_{\text{f}})$. We take $N \in \mathbb{N}^*$ such that $N := \max\{n \in \mathbb{N} / n\,\Delta t < t_{\text{f}}\}$, and we denote $t_n = n\,\Delta t$, for $n \in \{0, \ldots, N + 1\}$, so that $t_0 = 0$, and $t_{N+1} = t_{\text{f}}$.
We define the following finite-dimensional spaces :

$$\mathcal{H}_{\mathcal{T}} := \left\{\varphi \in C^0\left(\overline{\Omega}\right); \varphi_{|_K} \in \mathbb{P}_1, \ \forall K \in \mathcal{T}\right\} \subset H^1(\Omega),$$
$$\mathcal{H}^0_{\mathcal{T}} := \left\{\varphi \in \mathcal{H}_{\mathcal{T}};\, \varphi(\mathbf{x}_M) = 0, \quad \forall M \in \mathcal{M}^{\text{ext}}\right\}.$$

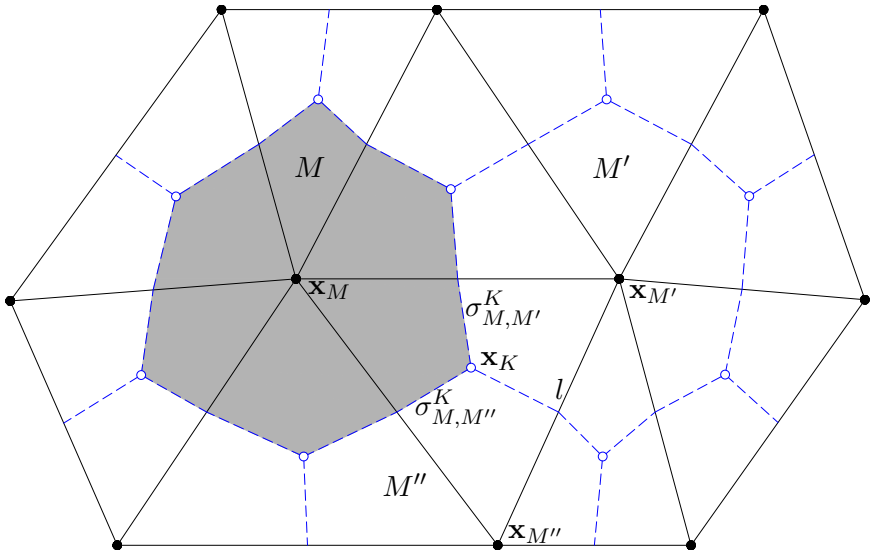


FIGURE 3.1 – Primal triangular mesh $\mathcal{T}$ and Donald dual mesh $\mathcal{M}$: control volumes, centers, interfaces.

The canonical basis of $\mathcal{H}_\mathcal{T}$ is spanned by the shape functions $(\varphi_M)_{M\in\mathcal{M}}$, such that $\varphi_M(\mathbf{x}_{M'}) = \delta_{M,M'}$ for all $M' \in \mathcal{M}$, δ being the Kronecker delta. The approximations in these spaces are conforming since $\mathcal{H}_\mathcal{T} \subset H^1(\Omega)$. We equip $\mathcal{H}_\mathcal{T}$ with the semi-norm

$$\|u_\mathcal{T}\|^2_{\mathcal{H}_\mathcal{T}} := \int_\Omega |\nabla u_\mathcal{T}|^2 \,\mathrm{d}\mathbf{x}, \qquad \forall u_\mathcal{T} \in \mathcal{H}_\mathcal{T},$$

which becomes a norm on $\mathcal{H}^0_\mathcal{T}$.

The classical finite elements $\mathbb{P}_1$ associated to the vertex $\mathbf{x}_{M_i}$ $(i = 1, \ldots, N_s)$, where N_s is the total number of vertices, is defined by

$$\varphi_{M_i}(\mathbf{x}_{M_j}) = \delta_{ij}, \text{ where } \varphi_{M_i} \text{ is continuous and piecewise } \mathbb{P}_1 \text{ per triangle.}$$

Let w^n_M be an expected approximation of $w(\mathbf{x}_M, t_n)$, where $w \equiv u$ or v. Thus, the discrete unknowns are denoted by $\{w^n_M / M \in \mathcal{M}, n \in \{0, \ldots, N+1\}\}$.

Definition 3.3. (Discrete functions). Using the values of $\left(u^{n+1}_M, v^{n+1}_M\right)$, $\forall M \in \mathcal{M}$ and $n \in \{0, \ldots, N\}$, we determine two approximate solutions by means of the *control volume finite element scheme* :

(i) A finite volume solution $(u_{\mathcal{M},\Delta t}, v_{\mathcal{M},\Delta t})$ defined as piecewise constant on the dual control volumes in space and piecewise constant in time, such that :

$$\begin{aligned}
&\left(u_{\mathcal{M},\Delta t}(\mathbf{x}, 0), v_{\mathcal{M},\Delta t}(\mathbf{x}, 0)\right) = \left(u^0_M, v^0_M\right) && \forall \mathbf{x} \in \overset{\circ}{M},\ M \in \mathcal{M},\\
&\left(u_{\mathcal{M},\Delta t}(\mathbf{x}, t), v_{\mathcal{M},\Delta t}(\mathbf{x}, t)\right) = \left(u^{n+1}_M, v^{n+1}_M\right) && \forall \mathbf{x} \in \overset{\circ}{M},\ M \in \mathcal{M}, \forall t \in (t_n, t_{n+1}],
\end{aligned}$$

where u^0_M (resp. v^0_M) represents the mean value of the function u_0 (resp. v_0). The discrete space of these functions namely *discrete control volumes space* is denoted by $\mathcal{X}_{\mathcal{M},\Delta t}$.

(ii) A finite element solution $v_{\mathcal{T},\Delta t}$ as a function continuous and piecewise $\mathbb{P}_1$ per triangle in space and piecewise constant in time, such that :

$$\begin{aligned}
&v_{\mathcal{T},\Delta t}(\mathbf{x}, 0) = v^0_\mathcal{T}(\mathbf{x}) && \forall \mathbf{x} \in \Omega,\\
&v_{\mathcal{T},\Delta t}(\mathbf{x}, t) = v^{n+1}_\mathcal{T}(\mathbf{x}) && \forall \mathbf{x} \in \Omega, \forall t \in (t_n, t_{n+1}],
\end{aligned}$$

where $v_{\mathcal{T}}^{n+1}(\mathbf{x}) := \sum_{M\in\mathcal{M}} v_M^{n+1}\varphi_M(\mathbf{x})$ and $v_{\mathcal{T}}^{0}(\mathbf{x}) := \sum_{M\in\mathcal{M}} v_M^{0}\varphi_M(\mathbf{x})$. The discrete space of these functions namely *discrete finite elements space* is denoted by $\mathcal{H}_{\mathcal{T},\Delta t}$.

The function A is nonlinear, we denote by $A_{\mathcal{T},\Delta t} = A_{\mathcal{T}}(u_{\mathcal{T},\Delta t})$ the corresponding finite element reconstruction in $\mathcal{H}_{\mathcal{T},\Delta t}$, and by $A_{\mathcal{M},\Delta t} = A(u_{\mathcal{M},\Delta t})$ the corresponding finite volume reconstruction in $\mathcal{X}_{\mathcal{M},\Delta t}$. Specifically, we have

$$A_{\mathcal{T}}(u_{\mathcal{T},\Delta t}(\mathbf{x},t)) = \sum_{M\in\mathcal{M}} A\left(u_M^{n+1}\right)\varphi_M(\mathbf{x}), \qquad \forall \mathbf{x}\in\Omega,\ \forall t\in(t_n,t_{n+1}],$$
$$A(u_{\mathcal{M},\Delta t}(\mathbf{x},t)) = A\left(u_M^{n+1}\right), \qquad \forall \mathbf{x}\in \overset{\circ}{M},\ M\in\mathcal{M},\ \forall t\in(t_n,t_{n+1}].$$

3.3.1 CVFE scheme for the modified Keller-Segel model

In order to define a discretization for system (3.1), we integrate the equations of system (3.1) over the set $M\times[t_n,t_{n+1}]$ with $M\in\mathcal{M}$, then we use the Green–Gauss formula as well as the implicit order one discretization in time, we get

$$\begin{aligned}
&\int_M (u(\mathbf{x},t_{n+1}) - u(\mathbf{x},t_n))\,\mathrm{d}\mathbf{x} - \sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla A(u)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \\
&\qquad + \sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \chi(u)\,\Lambda\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_M f(u)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \\
&\int_M (v(\mathbf{x},t_{n+1}) - v(\mathbf{x},t_n))\,\mathrm{d}\mathbf{x} - \sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma D\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t = \int_{t_n}^{t_{n+1}}\int_M g(u,v)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t,
\end{aligned} \tag{3.12}$$

where $\eta_{M,\sigma}$ is the unit normal vector outward to $\sigma\subset\partial M$, and $\mathrm{d}\sigma(\mathbf{x})$ is the Lebesgue measure on the edge σ.

We consider now an implicit Euler scheme in time, and thus the time evolution in the first equation of system (3.12) is approximated as

$$\begin{aligned}
\int_M (u(\mathbf{x},t_{n+1}) - u(\mathbf{x},t_n))\,\mathrm{d}\mathbf{x} &\approx \int_M (u_{\mathcal{M},\Delta t}(\mathbf{x},t_{n+1}) - u_{\mathcal{M},\Delta t}(\mathbf{x},t_n))\,\mathrm{d}\mathbf{x} \\
&= |M|\left(u_M^{n+1} - u_M^n\right).
\end{aligned}$$

Note that $f(u)$ is a nonlinear function. We denote by $f(u_{\mathcal{M},\Delta t})$ the corresponding piecewise constant reconstruction in $\mathcal{X}_{\mathcal{M},\Delta t}$, then the reaction term is approximated as

$$\int_{t_n}^{t_{n+1}}\int_M f(u(\mathbf{x},t))\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \approx \int_{t_n}^{t_{n+1}}\int_M f(u_{\mathcal{M},\Delta t}(\mathbf{x},t))\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = |M|\,\Delta t f\left(u_M^{n+1}\right).$$

On the other hand, we distinguish two kinds of approximation in space. The first consists of considering the finite element approach to handle the diffusion term, and the second consists of using a classical upstream finite volume approach.
Let us focus on the discretization of the diffusion term of the first equation of system (3.12), we have

$$\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla A(u)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx \Delta t\sum_{\sigma\subset\partial M}\int_\sigma \Lambda\nabla A_{\mathcal{T}}(u_{\mathcal{T},\Delta t}(\mathbf{x},t_{n+1}))\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x}). \tag{3.13}$$

The diffusion tensor $\Lambda(\mathbf{x})$ is taken constant per triangle, we denote by Λ_K the mean value of the function $\Lambda(\mathbf{x})$ over the triangle $K \in \mathcal{T}$, then one rewrites the right hand side of equation (3.13) as

$$\begin{aligned}\Delta t \sum_{K,K\cap M\neq\emptyset} \sum_{\sigma\subset\partial M\cap K} \Lambda_K \nabla A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}(\mathbf{x},t_{n+1})\right)_{|_K} \cdot \eta_{M,\sigma}\,|\sigma| \\ = \Delta t \sum_{K,K\cap M\neq\emptyset} \frac{1}{2} \nabla A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}(\mathbf{x},t_{n+1})\right)_{|_K} \cdot {}^t\Lambda_K \eta_{K,l}\,|l|\end{aligned} \tag{3.14}$$

where $l \in \mathcal{E}_K$ such that $M \cap l = \emptyset$, and $\eta_{K,l}$ denotes the unit normal vector outward to the edge l. For the transition between the first and the second line in approximation (3.14), we have used a geometric property (see Lemma A.3), that is

$$\sum_{\sigma\subset\partial M\cap K} \eta_{M,\sigma}\,|\sigma| = \frac{1}{2}\eta_{K,l}\,|l|\,, \qquad \forall K \in \mathcal{T} \text{ such that } K\cap M \neq \emptyset.$$

According to the definition of the approximate function $\nabla A_{\mathcal{T}}(u_{\mathcal{T},\Delta t})$, one has

$$\nabla A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}(\mathbf{x},t_{n+1})\right)_{|_K} = \nabla\left(\sum_{M\in\mathcal{M}} A\left(u_M^{n+1}\right)\varphi_M(\mathbf{x})\right)_{|_K} = \sum_{M\in\mathcal{M}} A\left(u_M^{n+1}\right)\nabla\varphi_M(\mathbf{x})_{|_K}. \tag{3.15}$$

Note that, the $\mathbb{P}_1$-finite element bases are expressed in barycentric coordinates, thus

$$\sum_{M',M'\cap K\neq\emptyset} \varphi_{M'}(\mathbf{x})_{|_K} = 1, \text{ and } \sum_{M',M'\cap K\neq\emptyset} \nabla\varphi_{M'}(\mathbf{x})_{|_K} = 0,$$

consequently, we have

$$\nabla A_{\mathcal{T}}\left(u_{\mathcal{T},\Delta t}(\mathbf{x},t_{n+1})\right)_{|_K} = \sum_{M',M'\cap K\neq\emptyset} \left(A\left(u_{M'}^{n+1}\right) - A\left(u_M^{n+1}\right)\right)\nabla\varphi_{M'|_K}. \tag{3.16}$$

We note that, for a given $K \in \mathcal{T}$, we have (see Appendix A)

$$\nabla\varphi_{M|_K} = \frac{-|l|}{2|K|}\eta_{K,l}, \quad \forall M \in \mathcal{M} \text{ such that } M\cap K \neq \emptyset. \tag{3.17}$$

Let us now introduce the transmissibility coefficient between M and M' defined, for all $K \in \mathcal{T}$ such that $K\cap M \neq \emptyset$ and $K\cap M' \neq \emptyset$, by

$$\Lambda_{M,M'}^K = -\int_K \Lambda(\mathbf{x})\,\nabla\varphi_M(\mathbf{x})\cdot\nabla\varphi_{M'}(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{3.18}$$

As a consequence of equations (3.17)–(3.18), one has

$$\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \Lambda\nabla A(u)\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\mathrm{d}t \approx \Delta t \sum_{K,M\cap K\neq\emptyset}\sum_{M',M'\cap K\neq\emptyset} \Lambda_{M,M'}^K\left(A\left(u_{M'}^{n+1}\right) - A\left(u_M^{n+1}\right)\right).$$

Next, we have to approximate the convective term in the first equation of system (3.12). For that, we consider a classical upstream finite volume scheme according to the normal component of the gradient of the chemoattractant v on the interfaces. So,

$$\begin{aligned}&\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \chi(u)\,\Lambda\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \\ &\quad\approx \Delta t \sum_{K,K\cap M\neq\emptyset}\sum_{\sigma\subset\partial M\cap K}\int_\sigma \Lambda(\mathbf{x})\,\chi\left(u_{\mathcal{M},\Delta t}(\mathbf{x},t_{n+1})\right)\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t_{n+1})\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x}).\end{aligned}$$

In order to approximate the convective flux on each interface, let us firstly introduce an example of approximation in the case where the function χ is nondecreasing. For that, we consider the interface $\sigma^K_{M,M'}$, and write

$$\int_{\sigma^K_{M,M'}} \chi\left(u_{\mathcal{M},\Delta t}\left(\mathbf{x},t_{n+1}\right)\right)\Lambda\left(\mathbf{x}\right)\nabla v_{\mathcal{T},\Delta t}\left(\mathbf{x},t_{n+1}\right)\cdot\eta_\sigma \,\mathrm{d}\sigma(\mathbf{x}) \approx \begin{cases} |\sigma^K_{M,M'}|\chi\left(u^{n+1}_M\right)\,\mathrm{d}V^K_{M,M'}, & \text{if } \mathrm{d}V^K_{M,M'}\geq 0,\\ |\sigma^K_{M,M'}|\chi\left(u^{n+1}_{M'}\right)\,\mathrm{d}V^K_{M,M'}, & \text{if } \mathrm{d}V^K_{M,M'}\leq 0, \end{cases}$$

where $\mathrm{d}V^K_{M,M'}$ represents an approximation of the gradient of v on the interface $\sigma^K_{M,M'}$, defined by

$$\mathrm{d}V^K_{M,M'} = \sum_{M'',M''\cap K\neq\emptyset} v^{n+1}_{M''}\nabla\varphi_{M''|_K}\cdot {}^t\Lambda_K\eta^K_{M,M'}. \tag{3.19}$$

Thus, the convective term is approximated as

$$\sum_{\sigma\subset\partial M}\int_{t_n}^{t_{n+1}}\int_\sigma \chi\left(u\right)\Lambda\nabla v\cdot\eta_{M,\sigma}\,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \approx \begin{cases} \Delta t\displaystyle\sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset}\chi(u^{n+1}_M)\Lambda^K_{M,M'}\left(v^{n+1}_{M'}-v^{n+1}_M\right), & \text{if } \mathrm{d}V^K_{M,M'}\geq 0,\\ \Delta t\displaystyle\sum_{K,M\cap K\neq\emptyset}\ \sum_{M',M'\cap K\neq\emptyset}\chi(u^{n+1}_{M'})\Lambda^K_{M,M'}\left(v^{n+1}_{M'}-v^{n+1}_M\right), & \text{if } \mathrm{d}V^K_{M,M'}\leq 0, \end{cases}$$

where $\Lambda^K_{M,M'}$ represents the transmissibility coefficient between M and M', given by equation (3.18).

In the general case, we use numerical convection flux functions G of arguments $(a,b,c)\in\mathbb{R}^3$ which are required to satisfy the following properties :

$$\begin{cases} \text{(a) } G\left(\cdot,b,c\right) \text{ is nondecreasing for all } b,c\in\mathbb{R},\\ \text{and } G\left(a,\cdot,c\right) \text{ is nonincreasing for all } a,c\in\mathbb{R};\\ \text{(b) } G\left(a,b,c\right)=-G\left(b,a,-c\right) \text{ for all } a,b,c\in\mathbb{R};\\ \text{(c) } G\left(a,a,c\right)=\chi\left(a\right)c \text{ for all } a,c\in\mathbb{R};\\ \text{(d) there exists } C>0 \text{ such that}\\ \quad \forall\, a,b,c\in\mathbb{R}\quad |G\left(a,b,c\right)|\leq C\left(|a|+|b|\right)|c|;\\ \text{(e) there exists a modulus of continuity } \omega:\mathbb{R}^+\to\mathbb{R}^+ \text{ such that}\\ \quad \forall\, a,b,a',b',c\in\mathbb{R}\quad |G\left(a,b,c\right)-G\left(a',b',c\right)|\leq|c|\,\omega\left(|a-a'|+|b-b'|\right). \end{cases} \tag{3.20}$$

In our context, one possibility to construct a numerical flux G satisfying conditions (3.20) is to split χ into the nondecreasing part $\chi_\uparrow$ and the nonincreasing part $\chi_\downarrow$:

$$\chi_\uparrow\left(z\right):=\int_0^z\left(\chi'\left(s\right)\right)^+\mathrm{d}s,\qquad \chi_\downarrow\left(z\right):=-\int_0^z\left(\chi'\left(s\right)\right)^-\mathrm{d}s.$$

Herein, $s^+=\max(s,0)$ and $s^-=\max(-s,0)$. Then we take

$$G\left(a,b;c\right)=c^+\Big(\chi_\uparrow(a)+\chi_\downarrow(b)\Big)-c^-\Big(\chi_\uparrow(b)+\chi_\downarrow(a)\Big). \tag{3.21}$$

Notice that in the case χ has a unique local (and global) maximum at the point $\tilde{u}\in(0,1)$, we have

$$\chi_\uparrow\left(z\right)=\chi\left(\min\{z,\tilde{u}\}\right)\ \text{ and }\ \chi_\downarrow\left(z\right)=\chi\left(\max\{z,\tilde{u}\}\right)-\chi\left(\tilde{u}\right).$$

For the discretization of the second equation of system (3.12), we define the transmissibility coefficient $D^K_{M,M'}$ by

$$D^K_{M,M'} = \int_K D(\mathbf{x}) \nabla\varphi_M(\mathbf{x}) \cdot \nabla\varphi_{M'}(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{3.22}$$

then we follow the same lines as for the discretization of the first equation.

We are now in a position to discretize problem (3.1)–(3.3). We denote by $\mathscr{D}$ a discretization of Q_{t_f}, which consists of a *primal finite element triangulation* $\mathcal{T}$ of Ω and its corresponding *Donald dual mesh* $\mathcal{M}$ and a time step $\Delta t > 0$.

A *control volume finite element scheme* for the discretization of problem (3.1)–(3.3) is given by the following set of equations : for all $M \in \mathcal{M}$,

$$u^0_M = \frac{1}{|M|}\int_M u_0(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad v^0_M = \frac{1}{|M|}\int_M v_0(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{3.23}$$

and for all $M \in \mathcal{M}$ and all $n \in \{0, \dots, N\}$,

$$\begin{aligned} |M| \frac{u^{n+1}_M - u^n_M}{\Delta t} - \sum_{K, M\cap K\neq\emptyset} \sum_{M', M'\cap K\neq\emptyset} \Lambda^K_{M,M'} \left(A\left(u^{n+1}_{M'}\right) - A\left(u^{n+1}_M\right)\right) \\ + \sum_{K, M\cap K\neq\emptyset} \sum_{M', M'\cap K\neq\emptyset} \left|\sigma^K_{M,M'}\right| G\left(u^{n+1}_M, u^{n+1}_{M'}; \mathrm{d}V^K_{M,M'}\right) = |M| f\left(u^{n+1}_M\right), \end{aligned} \tag{3.24}$$

$$|M| \frac{v^{n+1}_M - v^n_M}{\Delta t} - \sum_{K, M\cap K\neq\emptyset} \sum_{M', M'\cap K\neq\emptyset} D^K_{M,M'} \left(v^{n+1}_{M'} - v^{n+1}_M\right) = |M| g\left(u^n_M, v^{n+1}_M\right). \tag{3.25}$$

We recall that the unknowns are $U = \left(u^{n+1}_M\right)_{M\in\mathcal{M}}$ and $V = \left(v^{n+1}_M\right)_{M\in\mathcal{M}}$, $n \in \{0, \dots, N\}$, and that $\mathrm{d}V^K_{M,M'}$ is defined in equation (3.19), and the transmissibility coefficients $\Lambda^K_{M,M'}$ and $D^K_{M,M'}$ are given by equation (3.18) and equation (3.22) respectively. Notice that the discrete zero-flux boundary conditions are implicitly contained in equations (3.24)–(3.25). The contribution of $\partial\Omega \cap \partial M$ to the approximation of $\displaystyle\int_{\partial M} D\nabla v \cdot \eta \, \mathrm{d}\sigma(\mathbf{x})$ and $\displaystyle\int_{\partial M} \Lambda\left(\nabla A(u) - \chi(u)\nabla v\right) \cdot \eta \, \mathrm{d}\sigma(\mathbf{x})$ is zero, in compliance with equation (3.2).

In this chapter, we assume that all the transmissibility coefficients $\Lambda^K_{M,M'}$ and $D^K_{M,M'}$ are nonnegative :

$$\Lambda^K_{M,M'} \geq 0 \text{ and } D^K_{M,M'} \geq 0, \quad \forall M, M' \in \mathcal{M}, \forall K \in \mathcal{T}. \tag{3.26}$$

For the case where Λ and D are isotropic, and if all the triangles of $\mathcal{T}$ have only acute angles, the condition (3.26) holds. This was used for example in [18] to prove the convergence of the finite element approximation towards the renormalized solution of an elliptic equation.

3.3.2 Main result

Let $(\mathcal{T}_m)_{m\geq 1}$ be a sequence of triangulations of Ω such that

$$h_m = \max_{K\in\mathcal{T}_m} \mathrm{diam}(K) \to 0 \text{ as } m \to \infty.$$

We assume that there exists a constant $\kappa > 0$ such that

$$\kappa_{\mathcal{T}_m} \leq \kappa, \quad \forall m \geq 1.$$

As before, a sequence of dual meshes $(\mathcal{M}_m)_{m\geq 1}$ is given.

Let $(N_m)_m$ be an increasing sequence of integers, then we define the corresponding sequence of time steps $(\Delta t_m)_m$ such that $\Delta t_m \to 0$ as $m \to \infty$. The intention of this chapter is to prove the following main result.

Theorem 3.4. *Let* $(u_{\mathcal{M}_m,\Delta t_m}, v_{\mathcal{M}_m,\Delta t_m})_m$ *be a sequence of solutions to the scheme* (3.23)–(3.25)*, such that* $0 \leq u_{\mathcal{M}_m,\Delta t_m} \leq 1$ *and* $0 \leq v_{\mathcal{M}_m,\Delta t_m}$ *for almost everywhere in* Q_{t_f}*, then*

$$u_{\mathcal{M}_m,\Delta t_m} \to u \text{ and } v_{\mathcal{M}_m,\Delta t_m} \to v \quad \text{a.e. in } Q_{t_\mathrm{f}} \text{ as } m \to \infty,$$

where the couple (u, v) *is a weak solution to the system* (3.1)–(3.3) *in the sense of Definition 3.1.*

3.4 *A priori* analysis of discrete solutions

In this section, we prove the discrete maximum principle, then we establish the *a priori* estimates necessary to prove the existence of a solution to the discrete problem (3.23)–(3.25) and the convergence of the scheme towards the weak solution.

In the sequel, we denote by C a "generic" constant, which need not have the same value through the proofs.

3.4.1 Nonnegativity of $v_{\mathcal{M},\Delta t}$, confinement of $u_{\mathcal{M},\Delta t}$

We aim to prove the following lemma which is a basis to the analysis that we are going to perform.

Lemma 3.5. *Let* $(u_M^n, v_M^n)_{M\in\mathcal{M},\, n\in\{0,\dots,N+1\}}$ *be a solution of the CVFE scheme* (3.23)–(3.25)*. Under the nonnegativity of transmissibilty coefficients assumption* (3.26)*, we have for all* $M \in \mathcal{M}$*, and all* $n \in \{0,\dots,N+1\}$*,* $0 \leq u_M^n \leq 1$ *and* $0 \leq v_M^n$*. Moreover, there exists a positive constant* $\rho = \|v_0\|_\infty + \alpha t_\mathrm{f}$*, such that* $v_M^n \leq \rho$*, for all* $n \in \{0,\dots,N+1\}$*.*

Proof. Let us show by induction on n that for all $M \in \mathcal{M}$, $u_M^n \geq 0$. The claim is true for $n = 0$. We argue by induction that for all $M \in \mathcal{M}$, the claim is true up to order n. Consider a dual control volume M such that $u_M^{n+1} = \min\{u_{M'}^{n+1}\}_{M'\in\mathcal{M}}$, we want to show that $u_M^{n+1} \geq 0$. We consider equation (3.24) corresponding to the aforementioned dual control volume M, reorganize the summation over the edges and multiply it by $-\left(u_M^{n+1}\right)^-$ where for all real r, $r = r^+ - r^-$ with $r^+ = \max(r,0)$ and $r^- = \max(-r,0)$. This yields

$$\begin{aligned} -|M|\frac{u_M^{n+1}-u_M^n}{\Delta t}\left(u_M^{n+1}\right)^- &+ \sum_{\sigma_{M,M'}^K\subset\partial M} \Lambda_{M,M'}^K\left(A\left(u_{M'}^{n+1}\right)-A\left(u_M^{n+1}\right)\right)\left(u_M^{n+1}\right)^- \\ &- \sum_{\sigma_{M,M'}^K\subset\partial M} |\sigma_{M,M'}^K| G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V_{M,M'}^K\right)\left(u_M^{n+1}\right)^- = -f\left(u_M^{n+1}\right)\left(u_M^{n+1}\right)^-. \end{aligned} \tag{3.27}$$

Here, we use the extension by $f(0) \geq 0$ of the continuous function f for $u \leq 0$, then the right hand side of equation (3.27) is less or equal to zero.
The function A is nondecreasing then $A\left(u_{M'}^{n+1}\right) - A\left(u_M^{n+1}\right) \geq 0$ since $u_M^{n+1} \leq u_{M'}^{n+1}$. Furthermore, the assumption $\Lambda_{M,M'}^K \geq 0$ implies that

$$\sum_{\sigma_{M,M'}^K\subset\partial M} \Lambda_{M,M'}^K\left(A\left(u_{M'}^{n+1}\right)-A\left(u_M^{n+1}\right)\right)\left(u_M^{n+1}\right)^- \geq 0.$$

From the assumptions on the numerical flux G, the function G is nonincreasing with respect to the second variable $u_{M'}^{n+1}$, then using the extension by zero of the continuous function χ (recall that $\chi(u) = 0$ for $u \leq 0$), we get

$$\begin{aligned} G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V_{M,M'}^K\right)\left(u_M^{n+1}\right)^- &\leq G\left(u_M^{n+1}, u_M^{n+1}; \mathrm{d}V_{M,M'}^K\right)\left(u_M^{n+1}\right)^- \\ &= \mathrm{d}V_{M,M'}^K\,\chi\left(u_M^{n+1}\right)\left(u_M^{n+1}\right)^- = 0. \end{aligned}$$

Now, using the identity $u_M^{n+1} = \left(u_M^{n+1}\right)^+ - \left(u_M^{n+1}\right)^-$ and the nonnegativity of u_M^n, we deduce from equation (3.27) that $\left(u_M^{n+1}\right)^- = 0$. According to the choice of the dual control volume M, then $\min\{u_{M'}^{n+1}\}_{M'\in\mathcal{M}}$ is nonnegative ; this implies that

$$0 \le u_M^{n+1} \le u_{M'}^{n+1} \text{ for all } n \in \{0,\dots,N\} \text{ and all } M' \in \mathcal{M}.$$

In order to prove that $u_M^n \le 1$ for all $n \in \{0,\dots,N+1\}$ and all $M \in \mathcal{M}$, we argue by induction that for all $M \in \mathcal{M}$, the claim $u_M^n \le 1$ is true. We take the dual control volume M such that $u_M^{n+1} = \max\left(u_{M'}^{n+1}\right)_{M'\in\mathcal{M}}$, we want to show that $u_M^{n+1} \le 1$.
For the mentioned claim, we multiply equation (3.24) by $\left(u_M^{n+1}-1\right)^+$, one gets

$$\begin{aligned}|M|\,\frac{u_M^{n+1}-u_M^n}{\Delta t}\left(u_M^{n+1}-1\right)^+ - \sum_{\sigma_{M,M'}^K\subset\partial M}\Lambda_{M,M'}^K\left(A\left(u_{M'}^{n+1}\right)-A\left(u_M^{n+1}\right)\right)\left(u_M^{n+1}-1\right)^+ \\ + \sum_{\sigma_{M,M'}^K\subset\partial M}\left|\sigma_{M,M'}^K\right| G\left(u_M^{n+1},u_{M'}^{n+1};\mathrm{d}V_{M,M'}^K\right)\left(u_M^{n+1}-1\right)^+ = f\left(u_M^{n+1}\right)\left(u_M^{n+1}-1\right)^+.\end{aligned} \tag{3.28}$$

Since A is nondecreasing function, and $u_{M'}^{n+1} \le u_M^{n+1}$, we get $A\left(u_{M'}^{n+1}\right) - A\left(u_M^{n+1}\right) \le 0$. This implies that

$$-\sum_{\sigma_{M,M'}^K\subset\partial M}\Lambda_{M,M'}^K\left(A\left(u_{M'}^{n+1}\right)-A\left(u_M^{n+1}\right)\right)\left(u_M^{n+1}-1\right)^+ \ge 0. \tag{3.29}$$

Next, we use again the fact that the numerical flux function G is nonincreasing with respect to the second variable u_M^{n+1} and consistence (see (b) and (c) in properties (3.20)) to deduce, using the extension by zero of the continuous function χ for $u \ge 1$, that

$$\begin{aligned}G\left(u_M^{n+1},u_{M'}^{n+1};\mathrm{d}V_{M,M'}^K\right)\left(u_M^{n+1}-1\right)^+ &\ge G\left(u_M^{n+1},u_M^{n+1};\mathrm{d}V_{M,M'}^K\right)\left(u_M^{n+1}-1\right)^+ \\ &= \mathrm{d}V_{M,M'}^K\,\chi\left(u_M^{n+1}\right)\left(u_M^{n+1}-1\right)^+ = 0.\end{aligned} \tag{3.30}$$

Now, we rely on the extension by $f(1) \le 0$ on $[1,+\infty[$ of the production term source f in the right hand side of (3.28), thus $f\left(u_M^{n+1}\right)\left(u_M^{n+1}-1\right)^+ \le 0$.
Using the above estimates (3.29)–(3.30) into equation (3.28), one has

$$\begin{aligned}|M|\,\frac{u_M^{n+1}-u_M^n}{\Delta t}\left(u_M^{n+1}-1\right)^+ = \frac{|M|}{\Delta t}\Big(&\left(u_M^{n+1}-1\right)\left(u_M^{n+1}-1\right)^+ \\ &-\left(u_M^n-1\right)\left(u_M^{n+1}-1\right)^+\Big) \le 0.\end{aligned} \tag{3.31}$$

Using again the identity $\left(u_M^{n+1}-1\right) = \left(u_M^{n+1}-1\right)^+ - \left(u_M^{n+1}-1\right)^-$, and that $u_M^n \le 1$, one can deduce from estimate (3.31) that $\left(u_M^{n+1}-1\right)^+ = 0$. Consequently, we obtain

$$u_{M'}^{n+1} \le u_M^{n+1} \le 1 \text{ for all } n \in \{0,\dots,N\} \text{ and all } M' \in \mathcal{M}.$$

The proof of nonnegativity of v_M^n, $M \in \mathcal{M}$, $n \in \{0,\dots,N+1\}$, follows the same lines as in the proof for the nonnegativity of u_M^n, since $-g\left(u_M^n,v_M^{n+1}\right)\left(u_M^{n+1}\right)^- = -\alpha\,|M|\,u_M^n\left(v_M^{n+1}\right)^- + \beta\,|M|\,v_M^{n+1}\left(v_M^{n+1}\right)^- \le 0$. For simplicity, we do not provide it here.

Let us now focus on the last claim concerning the existence of a constant ρ such that $v_M^n \le \rho$ for all $M \in \mathcal{M}$ and all $n \in \{0,\dots,N+1\}$. We set $\rho_n := \|v_0\|_\infty + n\alpha\Delta t$, and suppose that

$v_M^n \leq \rho_n,\ \forall M \in \mathcal{M}$ (the claim holds for $n = 0$). We want to show that $v_M^{n+1} \leq \rho_{n+1}$, for that we take the dual control volume M such that $v_M^{n+1} = \max\{v_{M'}^{n+1}\}_{M' \in \mathcal{M}}$. Using scheme (3.25) and the fact that $\rho_{n+1} = \rho_n + \alpha\Delta t$, one has

$$|M| \frac{v_M^{n+1} - \rho_{n+1}}{\Delta t} + |M| \beta v_M^{n+1} - \sum_{K, M\cap K\neq\emptyset} \sum_{M', M'\cap K\neq\emptyset} D_{M,M'}^K \left(v_{M'}^{n+1} - v_M^{n+1}\right) = \alpha |M| \left(u_M^n - 1\right) + |M| \frac{v_M^n - \rho_n}{\Delta t}. \tag{3.32}$$

Multiplying equation (3.32) by $\left(v_M^{n+1} - \rho_{n+1}\right)^+$, one can deduce that $v_M^{n+1} \leq \rho_{n+1} \leq \rho$, for all $n \in \{0, \ldots, N\}$. This ends the proof of the lemma. □

3.4.2 Discrete *a priori* estimates

We derive in the next proposition, the main uniform estimates on the discrete gradient of the cell density $A(u)$ and the discrete gradient of the chemical concentration v. These estimates are necessary to obtain later results on the compactness as well as the existence of discrete solutions to the discrete problem

Proposition 3.6. *Let $(u_M^{n+1}, v_M^{n+1})_{M\in\mathcal{M}, n\in\{0,\ldots,N\}}$, be a solution of the control volume finite element scheme (3.23)–(3.25). Under assumption (3.11) and assumption (3.26), there exists a constant $C > 0$, depending only on Ω, t_f, $\|v_0\|_\infty$, $\kappa_\mathcal{T}$, α, f, and on the constant in (3.20)(d) such that*

$$\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} \Lambda_{M,M'}^K \left|A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right|^2 + \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} D_{M,M'}^K \left|v_M^{n+1} - v_{M'}^{n+1}\right|^2 \leq C, \tag{3.33}$$

consequently, for all $A_\mathcal{T}\left(u_\mathcal{T}^{n+1}\right) = \sum_{M\in\mathcal{M}} A\left(u_M^{n+1}\right)\varphi_M \in \mathcal{H}_\mathcal{T}$, and all $v_\mathcal{T}^{n+1} = \sum_{M\in\mathcal{M}} v_M^{n+1}\varphi_M \in \mathcal{H}_\mathcal{T}$,

$$\sum_{n=0}^{N} \Delta t \left\|v_\mathcal{T}^{n+1}\right\|_{\mathcal{H}_\mathcal{T}}^2 \leq C, \tag{3.34}$$

and

$$\sum_{n=0}^{N} \Delta t \left\|A_\mathcal{T}\left(u_\mathcal{T}^{n+1}\right)\right\|_{\mathcal{H}_\mathcal{T}}^2 \leq C. \tag{3.35}$$

Proof. To prove estimate (3.33), we multiply equation (3.24) (resp. equation (3.25)) by $A\left(u_M^{n+1}\right)$ (resp. by v_M^{n+1}), and perform a sum over $M \in \mathcal{M}$ and $n \in \{0, \ldots, N\}$. This yields

$$E_{1,1} + E_{1,2} + E_{1,3} = E_{1,4} \quad \text{and} \quad E_{2,1} + E_{2,2} = E_{2,3},$$

where

$$E_{1,1} = \sum_{n=0}^{N} \sum_{M\in\mathcal{M}} |M| \left(u_M^{n+1} - u_M^n\right) A\left(u_M^{n+1}\right), \quad E_{2,1} = \sum_{n=0}^{N} \sum_{M\in\mathcal{M}} |M| \left(v_M^{n+1} - v_M^n\right) v_M^{n+1},$$

$$E_{1,2} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} \Lambda_{M,M'}^K \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right) A\left(u_M^{n+1}\right),$$

$$E_{1,3} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} \left|\sigma_{M,M'}^K\right| G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V_{M,M'}^K\right) A\left(u_M^{n+1}\right),$$

$$E_{1,4} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} |M| \; f\left(u_M^{n+1}\right) A\left(u_M^{n+1}\right),$$

$$E_{2,2} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} D_{M,M'}^K \left(v_M^{n+1} - v_{M'}^{n+1}\right) v_M^{n+1},$$

$$E_{2,3} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} |M| \left(\alpha u_M^n - \beta v_M^{n+1}\right) v_M^{n+1}.$$

We consider the Lipschitz continuous nondecreasing function $\mathcal{B} : [0,1] \longrightarrow \mathbb{R}$ defined, for every $s \in \mathbb{R}$, by $\mathcal{B}(s) := \displaystyle\int_0^s A(r)\, \mathrm{d}r$; we have $\mathcal{B}''(s) = a(s) \geq 0$, so that $\mathcal{B}$ is convex.
From the convexity of $\mathcal{B}$ and of the function $s \longrightarrow \dfrac{1}{2}s^2$, we have the following inequalities

$$\forall a, b \in \mathbb{R}, \qquad (a-b)\, A(a) \geq \mathcal{B}(a) - \mathcal{B}(b), \qquad (a-b)\, a \geq \frac{1}{2}\left(a^2 - b^2\right).$$

Using these inequalities for the terms $E_{1,1}$ and $E_{2,1}$, we obtain

$$\begin{aligned} E_{1,1} &= \sum_{n=0}^{N} \sum_{M\in\mathcal{M}} |M| \left(u_M^{n+1} - u_M^n\right) A\left(u_M^{n+1}\right) \\ &\geq \sum_{n=0}^{N} \sum_{M\in\mathcal{M}} |M| \left(\mathcal{B}\left(u_M^{n+1}\right) - \mathcal{B}\left(u_M^n\right)\right) = \sum_{M\in\mathcal{M}} |M| \left(\mathcal{B}\left(u_M^{N+1}\right) - \mathcal{B}\left(u_M^0\right)\right), \\ E_{2,1} &= \sum_{n=0}^{N} \sum_{M\in\mathcal{M}} |M| \left(v_M^{n+1} - v_M^n\right) v_M^{n+1} \geq \frac{1}{2} \sum_{M\in\mathcal{M}} |M| \left(\left(v_M^{n+1}\right)^2 - \left(v_M^n\right)^2\right). \end{aligned}$$

Next, for the diffusive term $E_{1,2}$, we reorganize the sum over edges (discrete integration by parts). Then, we have

$$\begin{aligned} E_{1,2} &= \sum_{n=0}^{N} \Delta t \sum_{\sigma_{M,M'}^K \in \mathcal{L}} \Lambda_{M,M'}^K \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right)^2 \\ &= \frac{1}{2} \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} \Lambda_{M,M'}^K \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right)^2. \end{aligned}$$

We estimate the convective term $E_{1,3}$, and also gather by edges, one gets

$$\begin{aligned} E_{1,3} &= \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma^K_{M,M'}\subset\partial M} \left|\sigma^K_{M,M'}\right| G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right) A\left(u_M^{n+1}\right) \\ &= \sum_{n=0}^{N} \Delta t \sum_{\sigma^K_{M,M'}\in\mathcal{L}} \left|\sigma^K_{M,M'}\right| G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right) \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right). \end{aligned}$$

Using the definition of the function G, the assumption (3.20)(d), the boundedness of u_M^{n+1}, and applying the weighted Young inequality, one has

$$\begin{aligned} |E_{1,3}| &\le \sum_{n=0}^{N} \Delta t \sum_{\sigma^K_{M,M'}\in\mathcal{L}} \left|\sigma^K_{M,M'}\right| \left|G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right) \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right)\right| \\ &\le E^1_{1,3} + E^2_{1,3}, \end{aligned}$$

where

$$\begin{aligned} E^1_{1,3} &= \frac{1}{4}\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'} \left|A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right|^2, \\ E^2_{1,3} &= C\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma^K_{M,M'}\subset\partial M} \left|G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right) \sigma^K_{M,M'}\right|^2 \\ &\le C\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma^K_{M,M'}\subset\partial M} \left|\mathrm{d}V^K_{M,M'}\right|^2 \left|\sigma^K_{M,M'}\right|^2. \end{aligned}$$

On the other hand, using the definition (3.19) of $\mathrm{d}V^K_{M,M'}$ and the upper bound of the diffusion tensor Λ, one gets

$$\left|\mathrm{d}V^K_{M,M'}\right| \left|\sigma^K_{M,M'}\right| \le C \sum_{M'',M''\cap K\neq\emptyset} \left|v_{M''}^{n+1} - v_M^{n+1}\right| \left|\nabla\varphi_{M''|_K}\right| \left|\sigma^K_{M,M'}\right|.$$

Thanks to the shape regularity assumption (3.11), one can deduce that $\left|\nabla\varphi_{M''|_K}\right| \left|\sigma^K_{M,M'}\right| \le C$. As a consequence,

$$\begin{aligned} |E_{1,3}| \le \frac{1}{4}\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma^K_{M,M'}\subset\partial M} &\Lambda^K_{M,M'} \left|A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right|^2 \\ &+ C\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma^K_{M,M'}\subset\partial M} D^K_{M,M'} \left|v_{M'}^{n+1} - v_M^{n+1}\right|^2. \end{aligned}$$

The last estimation for the reactive term is given using definition (3.4) of g and the boundedness

of u_M^{n+1}, v_M^{n+1}, and f. Then

$$E_{2,3} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} |M| \left(\alpha u_M^n v_M^{n+1} - \beta \left(v_M^{n+1} \right)^2 \right) \leq \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \alpha |M| \, v_M^{n+1} \leq \alpha \rho t_{\mathrm{f}} |\Omega|.$$

$$E_{1,4} = \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} |M| \; f\left(u_M^{n+1}\right) A\left(u_M^{n+1}\right) \leq C t_{\mathrm{f}} |\Omega| \, .$$

Collecting the previous inequalities, one can deduce that there exists a constant $C > 0$, independent of h and Δt, such that

$$\sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} \Lambda_{M,M'}^K \left| A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right) \right|^2 + \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} D_{M,M'}^K \left| v_M^{n+1} - v_{M'}^{n+1} \right|^2 \leq C.$$

Let us focus on estimate (3.35), we denote by D_K, $K \in \mathcal{T}$ the mean value of the function D over the triangle K, then using the previous estimates as well as the assumptions on the diffusion tensor D, one has

$$\begin{aligned}
C &\geq \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{\sigma_{M,M'}^K \subset \partial M} D_{M,M'}^K \left| v_M^{n+1} - v_{M'}^{n+1} \right|^2 \\
&= 2 \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} v_M^{n+1} \sum_{\sigma_{M,M'}^K \subset \partial M} D_{M,M'}^K \left(v_M^{n+1} - v_{M'}^{n+1} \right) \\
&= 2 \sum_{n=0}^{N} \Delta t \sum_{M\in\mathcal{M}} \sum_{K\in\mathcal{T}} |K| \, v_M^{n+1} \nabla \varphi_{M|_K} \cdot {}^t D_K \sum_{M'\in\mathcal{M}} v_{M'}^{n+1} \nabla \varphi_{M'|_K} \\
&= \sum_{n=0}^{N} \Delta t \sum_{K\in\mathcal{T}} |K| \, D_K \left(\sum_{M\in\mathcal{M}} v_M^{n+1} \nabla \varphi_{M|_K} \right) \cdot \left(\sum_{M'\in\mathcal{M}} v_{M'}^{n+1} \nabla \varphi_{M'|_K} \right) \\
&= \sum_{n=0}^{N} \Delta t \sum_{K\in\mathcal{T}} \int_K D(\mathbf{x}) \nabla v_{\mathcal{T}}^{n+1} \cdot \nabla v_{\mathcal{T}}^{n+1} \mathrm{d}\mathbf{x} \geq 2 D_- \sum_{n=0}^{N} \Delta t \left\| v_{\mathcal{T}}^{n+1} \right\|_{\mathcal{H}_{\mathcal{T}}}^2 .
\end{aligned}$$

In the same manner and using estimate (3.33), one can establish estimate (3.34). This ends the proof of the **Proposition 3.6**. □

3.4.3 Existence of a discrete solution

The existence of a solution to the *control volume finite element scheme* will be obtained with the help of the following lemma proved in [55, 28, 75].

Lemma 3.7 (Application of Brouwer fixed-point theorem)**.** *Let $\mathcal{A}$ be a finite dimensional Hilbert space with inner product denoted by $[\cdot,\cdot]_{\mathcal{A}}$ and associated norm $\|\cdot\|_{\mathcal{A}}$. Let $\mathcal{P}$ be a continuous mapping from $\mathcal{A}$ into itself satisfying the following property : there exists $r > 0$ such that*

$$[\mathcal{P}(\xi), \xi]_{\mathcal{A}} > 0 \text{ for all } \xi \in \mathcal{A} \text{ with } \|\xi\|_{\mathcal{A}} = r.$$

Then, there exists $\xi \in \mathcal{A}$ *with* $\|\xi\|_{\mathcal{A}} \leq r$ *such that*

$$\mathcal{P}(\xi) = 0.$$

The existence of a discrete solution for the *nonlinear control volume finite element scheme* is given in the following proposition.

Proposition 3.8. *Under the shape regularity assumption* (3.11) *and the nonnegativity assumption* (3.26)*, there exists at least one solution* $\left(u_M^{n+1}, v_M^{n+1}\right)_{(M,n)\in\mathcal{M}\times[\![0\cdots N]\!]}$ *to the discrete problem* (3.23)–(3.25).

Proof. Denote by $U_h^n = (u_M^n)_{M\in\mathcal{M}}$ and $V_h^n = (v_M^n)_{M\in\mathcal{M}}$. We will show the existence of a discrete solution by induction on n. Assume that (u_h^n, v_h^n) exists and show the existence of $\left(u_h^{n+1}, v_h^{n+1}\right)$. Note that the discrete equation (3.25) is a standard time-implicit finite volume discretization of a uniformly parabolic equation, where the contribution of u in the right-hand side is discretized in an explicit way. Hence, The discrete equation (3.25) can be written as a finite dimensional linear system $\mathbf{A}V_h^{n+1} = \mathbf{B}$ with respect to the unknown V_h^{n+1}. The coefficients $(\mathbf{A}_{i,j})_{i,j\in\{1,\ldots,N_s\}}$ and $(\mathbf{B}_i)_{i\in\{1,\ldots,N_s\}}$ of the matrix $\mathbf{A}$ are given, for all $i \neq j \in \{1,\ldots,N_s\}$, by

$$\mathbf{A}_{i,i} = 1 + \beta\Delta t + \frac{\Delta t}{|M|} \sum_{K, M\cap K\neq\emptyset} \sum_{M', M'\cap K\neq\emptyset} D_{M,M'}^K,$$

$$\mathbf{A}_{i,j} = -\frac{\Delta t}{|M|} \sum_{K, M\cap K\neq\emptyset} \sum_{M', M'\cap K\neq\emptyset} D_{M,M'}^K,$$

$$\mathbf{B}_i = \alpha\Delta t u_M^n + v_M^n.$$

It is clear that the resulting matrix $\mathbf{A}$ of this system is symmetric and strictly diagonally dominant with nonnegative diagonal coefficients. As a consequence, $\mathbf{A}$ is invertible and definite positive. Thus the equation (3.25) admits a unique solution V_h^{n+1}.
Let us now prove the existence of a solution to the scheme (3.24). The nonlinear function A defined by equation (3.10) is nondecreasing, hence A in invertible. We can rewrite the scheme (3.24) in terms of ω_h^i with $U_h^i = A^{-1}\left(\omega_h^i\right)$, $i \in \{0,\cdots,N+1\}$. Assume that ω_h^n and V_h^{n+1} exist. We choose the component-wise inner product $[\cdot,\cdot]$ as the scalar product on $\mathbb{R}^{\mathrm{Card}(\mathcal{M})}$. We define the mapping $\mathbf{F}$, that associates to the vector $\mathcal{W} = \left(W_M^{n+1}\right)_{M\in\mathcal{M}}$, the following expression

$$\mathbf{F}(\mathcal{W}) = \Bigg(|M| \frac{A^{-1}\left(W_M^{n+1}\right) - A^{-1}(W_M^n)}{\Delta t} - \sum_{\sigma_{M,M'}^K \subset \partial M} \Lambda_{M,M'}^K \left(W_{M'}^{n+1} - W_M^{n+1}\right)$$
$$+ \sum_{\sigma_{M,M'}^K \subset \partial M} \left|\sigma_{M,M'}^K\right| G\left(A^{-1}\left(W_M^{n+1}\right), A^{-1}\left(W_{M'}^{n+1}\right); \mathrm{d}V_{M,M'}^K\right)$$
$$- |M| f\circ A^{-1}\left(W_M^{n+1}\right) \Bigg)_{M\in\mathcal{M}},$$

given by equation (3.24). Now, using Lemma 3.5 and estimates (3.33), one can deduce that

$$[\mathbf{F}(\mathcal{W}), \mathcal{W}] \geq C|\mathcal{W}|^2 - C'|\mathcal{W}| - C'', \qquad \text{for } |\mathcal{W}| \text{ large enough,}$$

where C, C', C'' are three nonnegative constants.
This implies, reasoning by contradiction and using Brouwer theorem (see Lemma 3.7), that, there exists $\mathcal{W} \in \mathbb{R}^{\mathrm{Card}(\mathcal{M})}$ such that

$$\mathbf{F}(\mathcal{W}) = 0.$$

Therefore, we obtain the existence of at least one solution to the scheme (3.24). □

3.5 Compactness estimates on discrete solutions

In this section, we derive estimates on differences of time and space translates necessary to prove the relative compactness property of the sequence of approximate solutions using Kolmogorov's theorem. Under the shape regularity assumption (3.11) and the nonnegativity of the transmissibility coefficients assumption (3.26), we give the time and space translate estimates for $A\left(u_{\mathcal{M}_h,\Delta t_h}\right)$ and $v_{\mathcal{M}_h,\Delta t_h}$ given by **Definition 3.3**.

3.5.1 Time translate estimate

Lemma 3.9. *Under assumption* (3.11) *and assumption* (3.26)*, there exists a positive constant* $C>0$ *depending on* Ω, t_{f}, α, u_0 *and* v_0 *(neither on* h *nor on* τ*) such that*

$$\iint_{\Omega\times(0,t_{\mathrm{f}}-\tau)} \left|w_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t+\tau\right)-w_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t\right)\right|^2 \mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C(\tau+\Delta t_h), \quad \text{for all } \tau\in(0,t_{\mathrm{f}}) \tag{3.36}$$

with $w_{\mathcal{M}_h,\Delta t_h}=A\left(u_{\mathcal{M}_h,\Delta t_h}\right)$, *or* $v_{\mathcal{M}_h,\Delta t_h}$.

Proof. Let $\tau\in(0,t_{\mathrm{f}})$ and $t\in(0,t_{\mathrm{f}}-\tau)$. We consider the quantity $\Upsilon\left(t\right)$ defined by

$$\Upsilon\left(t\right):=\int_{\Omega}\left|A\left(u_{\mathcal{M}_h,\Delta t_h}\right)\left(\mathbf{x},t+\tau\right)-A\left(u_{\mathcal{M}_h,\Delta t_h}\right)\left(\mathbf{x},t\right)\right|^2\mathrm{d}\mathbf{x}.$$

Set $n_0\left(t\right)=\left[t/\Delta t_h\right]$ and $n_1\left(t\right)=\left[\left(t+\tau\right)/\Delta t_h\right]$, where $\left[\mathbf{x}\right]=n$ for $\mathbf{x}\in[n,n+1)$, $n\in\mathbb{N}$. The function A is nondecreasing ; then using the mean value theorem, one gets the following inequality

$$\iint_{\Omega\times(0,t_{\mathrm{f}}-\tau)}\left|A\left(u_{\mathcal{M}_h,\Delta t_h}\right)\left(\mathbf{x},t+\tau\right)-A\left(u_{\mathcal{M}_h,\Delta t_h}\right)\left(\mathbf{x},t\right)\right|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t\leq C\int_0^{t_{\mathrm{f}}-\tau}\Upsilon(t)\,\mathrm{d}t,$$

where, for almost every $t\in(0,t_{\mathrm{f}}-\tau)$,

$$\Upsilon\left(t\right)=\sum_{M\in\mathcal{M}_h}|M|\left(A\left(u_M^{n_1(t)}\right)-A\left(u_M^{n_0(t)}\right)\right)\left(u_M^{n_1(t)}-u_M^{n_0(t)}\right).$$

Note that the function $\Upsilon\left(t\right)$ may be written as

$$\Upsilon\left(t\right)=\sum_{M\in\mathcal{M}_h}\left(A\left(u_M^{n_1(t)}\right)-A\left(u_M^{n_0(t)}\right)\right)\sum_{n=0}^{N-1}\chi_n\left(t,t+\tau\right)|M|\left(u_M^{n+1}-u_M^n\right), \tag{3.37}$$

where, χ_n is the characteristic function defined by

$$\chi_n\left(t,t+\tau\right)=\begin{cases}1 & \text{if } \left(n+1\right)\Delta t_h\in\left(t,t+\tau\right],\\ 0 & \text{if } \left(n+1\right)\Delta t_h\notin\left(t,t+\tau\right].\end{cases}$$

In equation (3.37), the order of the summation between n and M is changed and the scheme (3.24) is used. Hence,

$$\begin{aligned}\Upsilon\left(t\right)=\Delta t_h\sum_{n=0}^{N-1}\chi_n\left(t,t+\tau\right)\sum_{M\in\mathcal{M}_h}\left(A\left(u_M^{n_1(t)}\right)-A\left(u_M^{n_0(t)}\right)\right)\times\\ \sum_{\sigma^K_{M,M'}\subset\partial M}\left(\Lambda^K_{M,M'}\left(A\left(u_{M'}^{n+1}\right)-A\left(u_M^{n+1}\right)\right)-\left|\sigma^K_{M,M'}\right|G\left(u_M^{n+1},u_{M'}^{n+1};\mathrm{d}V^K_{M,M'}\right)\right)\\ +\Delta t_h\sum_{n=0}^{N-1}\chi_n\left(t,t+\tau\right)\sum_{M\in\mathcal{M}_h}\left(A\left(u_M^{n_1(t)}\right)-A\left(u_M^{n_0(t)}\right)\right)\times|M|\,f\left(u_M^{n+1}\right).\end{aligned}$$

We write $\Upsilon(t) = \Delta t_h \sum_{n=0}^{N-1} \chi_n(t, t+\tau)\left(\Upsilon_1(t) + \Upsilon_2(t) + \Upsilon_3(t)\right)$, where

$$\Upsilon_1(t) := \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'}\left(A\left(u_M^{n_1(t)}\right) - A\left(u_M^{n_0(t)}\right)\right)\left(A\left(u_{M'}^{n+1}\right) - A\left(u_M^{n+1}\right)\right),$$

$$\Upsilon_2(t) := \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \left|\sigma^K_{M,M'}\right|\left(A\left(u_M^{n_0(t)}\right) - A\left(u_M^{n_1(t)}\right)\right) G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right),$$

$$\Upsilon_3(t) := \sum_{M\in\mathcal{M}_h} |M|\left(A\left(u_M^{n_1(t)}\right) - A\left(u_M^{n_0(t)}\right)\right) f\left(u_M^{n+1}\right).$$

It is easy to see that

$$\sum_{n=0}^{N-1} \Delta t_h \int_0^{t_\mathrm{f}-\tau} \chi_n(t, t+\tau)\,\Upsilon_3(t)\,\mathrm{d}t \le C(\tau + \Delta t_h).$$

For the first term, note that gathering by edges, using the basic triangle inequality, one has

$$\begin{aligned}\Upsilon_1(t) \le \frac{1}{2}\Bigg(&\sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'}\left(A\left(u_{M'}^{n+1}\right) - A\left(u_M^{n+1}\right)\right)^2 \\ &+ \frac{1}{2}\sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'}\left|A\left(u_M^{n_1(t)}\right) - A\left(u_{M'}^{n_1(t)}\right)\right|^2 \\ &+ \frac{1}{2}\sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'}\left|A\left(u_M^{n_0(t)}\right) - A\left(u_{M'}^{n_0(t)}\right)\right|^2\Bigg).\end{aligned}$$

Using the estimates (3.33), this implies that there exists a constant $C > 0$ independent of τ and h, such that $\sum_{n=0}^{N-1} \Delta t_h \int_0^{t_\mathrm{f}-\tau} \chi_n(t, t+\tau)\,\Upsilon_1(t)\,\mathrm{d}t \le C(\tau + \Delta t_h)$. Finally, applying the previous arguments, gathering by edges, and using each of the definition of G and the assumptions on it, we get

$$\begin{aligned}\Upsilon_2(t)\,dt \le \frac{C}{2}\Bigg(&\sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \left(\left|A\left(u_M^{n_1(t)}\right) - A\left(u_{M'}^{n_1(t)}\right)\right|^2 + \left|v_M^{n+1} - v_{M'}^{n+1}\right|^2\right) \\ &+ \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \left(\left|A\left(u_M^{n_0(t)}\right) - A\left(u_{M'}^{n_0(t)}\right)\right|^2 + \left|v_M^{n+1} - v_{M'}^{n+1}\right|^2\right)\Bigg).\end{aligned}$$

We use estimate (3.33) to deduce that

$$\sum_{n=0}^{N-1} \Delta t_h \int_0^{t_\mathrm{f}-\tau} \chi_n(t, t+\tau)\,\Upsilon_2(t)\,\mathrm{d}t \le C(\tau + \Delta t_h).$$

Consequently, we obtain

$$\int_0^{t_{\mathrm{f}}-\tau} \Upsilon(t)\,\mathrm{d}t \le \sum_{n=0}^{N-1} \Delta t_h \int_0^{t_{\mathrm{f}}-\tau} \chi_n(t,t+\tau)\left(\Upsilon_1(t)+\Upsilon_3(t)\right)\mathrm{d}t + \sum_{n=0}^{N-1} \Delta t_h \int_0^{t_{\mathrm{f}}-\tau} \chi_n(t,t+\tau)\,\Upsilon_2(t)\,\mathrm{d}t \le C(\tau+\Delta t_h),$$

for some constant C independent of τ and h. The proof of (3.36) for $w_{\mathcal{M}_h,\Delta t_h} = v_{\mathcal{M}_h,\Delta t_h}$ follows in a similar manner. This concludes the proof of the lemma. □

We now extend by zero the functions $w_{\mathcal{M}_h,\Delta t_h}$ and $w_{\mathcal{T}_h,\Delta t_h}$ outside of $Q_{t_{\mathrm{f}}}$ and give the time translate estimate over $\mathbb{R}^3$ for $w_{\mathcal{M}_h,\Delta t_h}$. Indeed, there exists a constant $C>0$ independent of h and τ such that

$$\int_{\mathbb{R}}\int_{\mathbb{R}^2} |w_{\mathcal{M}_h,\Delta t_h}(t+\tau,\mathbf{x}) - w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)|^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le C(\tau+\Delta t_h), \quad \text{for all } \tau\in(0,t_{\mathrm{f}}).$$

Proof. Using the extension by zero of $w_{\mathcal{M}_h,\Delta t_h}$ outside of $Q_{t_{\mathrm{f}}}$, one has

$$\begin{aligned}&\int_{\mathbb{R}}\int_{\mathbb{R}^2} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t+\tau) - w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)|^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &= \iint_{Q_{t_{\mathrm{f}}-\tau}} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t+\tau) - w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)|^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{t_{\mathrm{f}}-\tau}^{t_{\mathrm{f}}}\int_{\Omega} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)|^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.\end{aligned}$$

One can deduce the proof using estimate (3.36) and the L^∞ bound of the function $w_{\mathcal{M}_h,\Delta t_h}$. □

We give now the space translate estimate on the approximate solutions.

3.5.2 Space translate estimate

Lemma 3.10. *Under assumptions (3.11) and (3.26), there exists a positive constant $C>0$ depending on Ω, t_{f}, α, u_0 and v_0 (neither on h nor on τ) such that*

$$\int_0^{t_{\mathrm{f}}}\int_{\mathbb{R}^2} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}+\mathbf{y},t) - w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le C(|\mathbf{y}|+h), \quad \textit{for all } \mathbf{y}\in\mathbb{R}^2 \tag{3.38}$$

with $w_{\mathcal{M}_h,\Delta t_h} = A(u_{\mathcal{M}_h,\Delta t_h})$, or $v_{\mathcal{M}_h,\Delta t_h}$.

Proof. We follow the same proof used in [16]. We first extend by zero the function $w_{\mathcal{T}_h,\Delta t_h}$ outside of $Q_{t_{\mathrm{f}}}$, and then we prove that

$$\int_0^{t_{\mathrm{f}}}\int_{\mathbb{R}^2} |w_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}+\mathbf{y},t) - w_{\mathcal{T}_h,\Delta t_h}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le C|\mathbf{y}|, \quad \text{for all } y\in\mathbb{R}^2. \tag{3.39}$$

Thanks to Proposition 3.6, and since $Q_{t_{\mathrm{f}}}$ is of finite measure, then Hölder inequality yields

$$\|\nabla w_{\mathcal{T}_h,\Delta t_h}\|_{\left(L^1(Q_{t_{\mathrm{f}}})\right)^2} \le |Q_{t_{\mathrm{f}}}|^{\frac{1}{2}}\,\|\nabla w_{\mathcal{T}_h,\Delta t_h}\|^{\frac{1}{2}}_{\left(L^2(Q_{t_{\mathrm{f}}})\right)^2} \le C. \tag{3.40}$$

Since A is uniformly bounded thanks to the boundedness of $u_{\mathcal{T}_h,\Delta t_h}$ and the uniform continuity of A, its extension to the whole $\mathbb{R}^3$, still denoted by $w_{\mathcal{T}_h,\Delta t_h}$, belongs to $L^\infty\cap \mathrm{BV}\left(\mathbb{R}^3\right)$ and satisfies

$$\mathrm{TV}\left(w_{\mathcal{T}_h,\Delta t_h}\right) \le \|\nabla w_{\mathcal{T}_h,\Delta t_h}\|_{\left(L^1(Q_{t_{\mathrm{f}}})\right)^2} + \|w\|_\infty\left(t_{\mathrm{f}}|\partial\Omega| + 2|\Omega|\right) < \infty, \tag{3.41}$$

where $|\partial\Omega|$ represents the length of $\partial\Omega$. The estimate (3.39) is a classical consequence of inequality (3.41) (see e.g. [13, 12]).
Now, using the extension by zero of $w_{\mathcal{M}_h,\Delta t_h}$ outside of Q_{t_f} as well as the triangle inequality, we get

$$\int_0^{t_\mathrm{f}}\int_{\mathbb{R}^2} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}+\mathbf{y},t) - w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le A + B,$$

where

$$A = \int_0^{t_\mathrm{f}}\int_{\mathbb{R}^2} |w_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}+\mathbf{y},t) - w_{\mathcal{T}_h,\Delta t_h}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$
$$B = 2\iint_{Q_{t_\mathrm{f}}} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t) - w_{\mathcal{T}_h,\Delta t_h}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

One can conclude the proof using estimate (3.39) and the following estimate (resulting from a straightforward generalization of Lemma A.5 given in Appendix A or using Lemma 3.11)

$$\iint_{Q_{t_\mathrm{f}}} |w_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t) - w_{\mathcal{T}_h,\Delta t_h}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le Ch. \tag{3.42}$$

This ends the proof of the lemma. □

3.6 Convergence of the CVFE scheme

We can prove the main result of this section. Specifically, we have the following lemmas.

Lemma 3.11. *The sequences* $\left(A\left(u_{\mathcal{M}_h,\Delta t_h}\right) - A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right)_h$ *and* $\left(v_{\mathcal{M}_h,\Delta t_h} - v_{\mathcal{T}_h,\Delta t_h}\right)_h$ *converge strongly to zero in* $L^2\left(Q_{t_\mathrm{f}}\right)$ *as* $h \to 0$.

Proof. Using the definition of the basis functions of the finite dimensional space $\mathcal{H}_{\mathcal{T}_h}$, we have for all $M \in \mathcal{M}_h$ and all $K \in \mathcal{T}_h$ such that $M \cap K \neq \emptyset$

$$\begin{aligned}\left|A\left(u_{\mathcal{M}_h,\Delta t_h}\right) - A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right|^2 &= \left|A\left(u_{\mathcal{M}_h,\Delta t_h}\right)(\mathbf{x}_M,t_{n+1}) - A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)(\mathbf{x},t_{n+1})\right|^2 \\ &= \left|\nabla A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)(\mathbf{x},t_{n+1})\cdot(\mathbf{x}_M-\mathbf{x})\right|^2, \quad \forall \mathbf{x}\in K\cap M\end{aligned}$$

where $\mathbf{x}_M$ is the center of the dual control volume $M \in \mathcal{M}_h$.
Using estimate (3.35), one obtains

$$\begin{aligned}&\left\|A\left(u_{\mathcal{M}_h,\Delta t_h}\right) - A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right\|^2_{L^2\left(Q_{t_\mathrm{f}}\right)} \\ &\qquad= \sum_{n=0}^{N}\Delta t_h \sum_{K\in\mathcal{T}_h}\sum_{M,M\cap K\neq\emptyset}\int_{K\cap M}\left|\nabla A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)(\mathbf{x},t_{n+1})\cdot(\mathbf{x}_M-\mathbf{x})\right|^2\mathrm{d}\mathbf{x} \\ &\qquad\le h^2\sum_{n=0}^{N}\Delta t_h \sum_{K\in\mathcal{T}_h}\sum_{M,M\cap K\neq\emptyset}|K\cap M|\left|\nabla A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h}^{n+1}\right)_{|K}\right|^2 \\ &\qquad\le h^2\sum_{n=0}^{N}\Delta t_h \left\|A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h}^{n+1}\right)\right\|^2_{\mathcal{H}_{\mathcal{T}_h}} \le Ch^2.\end{aligned}$$

As a consequence, we have $\left\|A\left(u_{\mathcal{M}_h,\Delta t_h}\right) - A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right\|_{L^2\left(Q_{t_\mathrm{f}}\right)} \longrightarrow 0$ as $h \to 0$. In the same manner, we prove that $\left\|v_{\mathcal{M}_h,\Delta t_h} - v_{\mathcal{T}_h,\Delta t_h}\right\|_{L^2\left(Q_{t_\mathrm{f}}\right)} \longrightarrow 0$ as $h \to 0$. □

Lemma 3.12. *(Convergence of the scheme). Under the shape regularity assumption (3.11) and the nonnegativity of the transmissibility coefficients assumption (3.26), there exists a sequence $(h_m)_{m\in\mathbb{N}}$, $h_m \to 0$ as $m \to \infty$, and functions u, v defined in Q_{t_f} such that $0 \le u \le 1$, both $A(u)$ and v belong to $L^2(0,T;H^1(\Omega))$, and*

$$A_{h_m}\left(u_{h_m}\right) \to A\left(u\right) \text{ and } v_{h_m} \to v \text{ a.e. in } Q_{t_\mathrm{f}} \text{ and strongly in } L^p(Q_{t_\mathrm{f}}) \text{ for all } p < +\infty$$

Proof. Let us set $\tilde{A}_{\mathcal{M}_h,\Delta t_h} := A\left(u_{\mathcal{M}_h,\Delta t_h}\right)$ in Q_{t_f} and $\tilde{A}_{\mathcal{M}_h,\Delta t_h} := 0$ in $\mathbb{R}^3 \setminus Q_{t_\mathrm{f}}$. Thanks to Proposition 3.6 and Lemma 3.5, one has $\left(\tilde{A}_{\mathcal{M}_h,\Delta t_h}\right) \subset L^\infty\left(\mathbb{R}^3\right) \cap L^2\left(\mathbb{R}^3\right)$. In order to verify the assumptions of Kolmogorov's compactness criterion, see [31, theorem 3.9, p :93], we note that the following inequality is verified for any $\eta \in \mathbb{R}^2$ and $\tau \in \mathbb{R}$,

$$\begin{aligned}\left\|\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot+\eta,\cdot+\tau\right)-\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)} &\le \left\|\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot+\eta,\cdot\right)-\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)} \\ &+ \left\|\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot+\tau\right)-\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)}.\end{aligned}$$

Using estimates (3.36) and (3.38), one has $\left\|\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot+\eta,\cdot+\tau\right)-\tilde{A}_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)} \longrightarrow 0$, as $\eta \to 0$ and $\tau \to 0$. This yields the compactness of the sequence $\left(\tilde{A}_{\mathcal{M}_h,\Delta t_h}\right)$ in $L^1\left(\Omega\right)$.
Thus, there exists a subsequence, still denoted by $\left(\tilde{A}_{\mathcal{M}_h,\Delta t_h}\right)$, and there exists $A^* \in L^1\left(Q_{t_\mathrm{f}}\right)$ such that

$$A\left(u_{\mathcal{M}_h,\Delta t_h}\right) \longrightarrow A^* \text{ strongly in } L^1\left(Q_{t_\mathrm{f}}\right).$$

Furthermore, as A is continuous and strictly monotone, there exists a unique u such that $A(u) = A^*$.
Since A^{-1} is well defined and continuous, then applying the L^∞ bound on $u_{\mathcal{M}_h,\Delta t_h}$ and the dominated convergence theorem to $u_{\mathcal{M}_h,\Delta t_h} = A^{-1}\left(A\left(u_{\mathcal{M}_h,\Delta t_h}\right)\right)$, we get

$$u_{\mathcal{M}_h,\Delta t_h} \longrightarrow u \text{ a.e. in } Q_{t_\mathrm{f}} \text{ and strongly in } L^p\left(Q_{t_\mathrm{f}}\right) \text{ for } p < +\infty.$$

According to Lemma 3.11, the sequences $\left(A\left(u_{\mathcal{M}_h,\Delta t_h}\right)\right)_{h,\Delta t_h}$ and $\left(A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right)_{h,\Delta t_h}$ have the same limit, as a consequence

$$A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right) \longrightarrow A(u) \text{ strongly in } L^2(Q_{t_\mathrm{f}}) \text{ and a.e. in } Q_{t_\mathrm{f}}.$$

Similarly, translate estimates (3.36)–(3.38), the L^∞ bound on $v_{h,\Delta t_h}$ in Lemma 3.5, and Lemma 3.11 ensure that, up to extraction of a subsequence,

$$v_{\mathcal{T}_h,\Delta t_h} \to v \text{ a.e. in } Q_{t_\mathrm{f}} \text{ and strongly in } L^p\left(Q_{t_\mathrm{f}}\right) \text{ for } p < +\infty.$$

It remains to show that $A\left(u\right)$ and $v \in L^2\left(0,T;H^1\left(\Omega\right)\right)$. Indeed, using estimate (3.33), one gets that the sequence $\nabla A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right) \in \left(L^2\left(Q_{t_\mathrm{f}}\right)\right)^2$. It follows that $\left(A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right)$ is bounded in $L^2\left(0,T;H^1\left(\Omega\right)\right)$ since $A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)$ is uniformly bounded in $L^2\left(Q_{t_\mathrm{f}}\right)$ due to the boundedness of the function A. Therefore, the sequence $\left(A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right)$ converges weakly, up to an unlabeled subsequence, to a function $A^\star$ in $L^2\left(0,T;H^1\left(\Omega\right)\right)$. The sequence $\left(A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)\right)$ converges strongly in $L^2\left(Q_{t_\mathrm{f}}\right)$ to $A\left(u\right)$, one can deduce that $A\left(u\right) = A^\star \in L^2\left(0,T;H^1\left(\Omega\right)\right)$. In the same manner, we obtain that $v \in L^2\left(0,T;H^1\left(\Omega\right)\right)$. This ends the proof of the lemma. □

It remains to be shown that the limit functions u and v constitute a weak solution of the continuous system. For this, let $\psi \in \mathcal{D}([0,T) \times \overline{\Omega})$ be a test function and denote by $\psi_M^n :=$

$\psi(\mathbf{x}_M, t_n)$ for all $M \in \mathcal{M}_h$ and $n \in \{0, \ldots, N+1\}$.
Multiply equation (3.24) by $\Delta t_h\, \psi_M^{n+1}$ and sum up over $M \in \mathcal{M}_h$ and $n \in \{0, \ldots, N\}$. One has

$$\begin{aligned}
&\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} |M| \left(u_M^{n+1} - u_M^n\right)\psi_M^{n+1} \\
&\qquad + \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'} \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right)\psi_M^{n+1} \\
&\qquad + \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right) \left|\sigma^K_{M,M'}\right| \psi_M^{n+1} \\
&\qquad\qquad = \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} |M|\, f(u_M^{n+1})\psi_M^{n+1},
\end{aligned}$$

that is,

$$S_1^h + S_2^h + S_3^h = S_4^h$$

where

$$\begin{aligned}
S_1^h &:= \sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} |M| \left(u_M^{n+1} - u_M^n\right)\psi_M^{n+1}, \quad S_4^h := \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} |M|\, f(u_M^{n+1})\psi_M^{n+1}, \\
S_2^h &:= \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} G\left(u_M^{n+1}, u_{M'}^{n+1}; \mathrm{d}V^K_{M,M'}\right) \psi_M^{n+1} \left|\sigma^K_{M,M'}\right|, \\
S_3^h &:= \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{\sigma^K_{M,M'}\subset\partial M} \Lambda^K_{M,M'} \left(A\left(u_M^{n+1}\right) - A\left(u_{M'}^{n+1}\right)\right)\psi_M^{n+1}.
\end{aligned}$$

Perform a summation by parts in time and keep in mind that $\psi_M^{N+1} = 0$ for all $M \in \mathcal{M}_h$, we obtain

$$\begin{aligned}
S_1^h &= \sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} |M|\, u_M^{n+1}\psi_M^{n+1} - \sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} |M|\, u_M^n \psi_M^{n+1} \\
&= -\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} \Delta t_h\, |M|\, u_M^n \frac{\psi_M^{n+1} - \psi_M^n}{\Delta t_h} - \sum_{M\in\mathcal{M}} |M|\, u_M^0 \psi_M^0 \\
&= -\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} \int_{t_n}^{t_{n+1}} \int_M u_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)\, \partial_t\psi(\mathbf{x}_M, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\
&\qquad\qquad - \sum_{M\in\mathcal{M}} \int_M u_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, 0)\, \psi(\mathbf{x}_M, 0) \mathrm{d}\mathbf{x}.
\end{aligned}$$

Taking into account the assumptions on the data and using the Lebesgue theorem, it follows that

$$S_1^h \xrightarrow[h,\Delta t_h\to 0]{} -\int_0^{t_\mathrm{f}} \int_\Omega u\, \partial_t\psi \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_\Omega u_0 \psi(\cdot, 0)\, \mathrm{d}\mathbf{x}.$$

On the other hand, for the convergence of the third term S_3^h, we note that

$$
\begin{aligned}
S_3^h &= \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{\sigma_{M,M'}^K\subset\partial M} \left(\Lambda_{M,M'}^K A\left(u_M^{n+1}\right)\psi_M^{n+1} - \Lambda_{M,M'}^K A\left(u_{M'}^{n+1}\right)\psi_M^{n+1}\right)\\
&= -\sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} A\left(u_M^{n+1}\right) \sum_{\sigma_{M,M'}^K\subset\partial M} \Lambda_{M,M'}^K\left(\psi_{M'}^{n+1} - \psi_M^{n+1}\right)\\
&= -\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} \Delta t_h A\left(u_M^{n+1}\right) \sum_{K\cap M\neq\emptyset} \sum_{\sigma_{M,M'}^K\subset\partial M\cap K} \nabla\psi^{n+1}{}_{|K}\cdot {}^t\Lambda_K \eta_{M,\sigma}\,|\sigma|\\
&= -\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} \Delta t_h A\left(u_M^{n+1}\right) \sum_{\sigma_{M,M'}^K\subset\partial M} \nabla\psi^{n+1}{}_{|K}\cdot {}^t\Lambda_K \eta_{M,\sigma}\,|\sigma|\\
&= -\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} \Delta t_h A\left(u_M^{n+1}\right) \int_{\partial M} \Lambda\nabla\psi^{n+1}\cdot\eta\,\mathrm{d}\sigma(\mathbf{x})\\
&= -\sum_{n=0}^{N} \sum_{M\in\mathcal{M}_h} \int_{t_n}^{t_{n+1}}\!\!\int_M A\left(u_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)\right)\mathrm{div}\,\left(\Lambda\nabla\psi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\xrightarrow[h,\Delta t_h\to 0]{} -\iint_{Q_{t_\mathrm{f}}} A(u)\,\mathrm{div}\,\left(\Lambda\nabla\psi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t = \iint_{Q_{t_\mathrm{f}}} \nabla A(u)\cdot\Lambda\nabla\psi\mathrm{d}\mathbf{x}\,\mathrm{d}t.
\end{aligned}
$$

It remains to show that

$$
\lim_{h,\Delta t_h\to\infty} S_2^h = -\int_0^{t_\mathrm{f}}\!\!\int_\Omega \Lambda(\mathbf{x})\,\chi(u)\nabla v\cdot\nabla\psi\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{3.43}
$$

For the convergence of S_2^h, we note that gathering by edges (thanks to the consistency of the fluxes, see (3.20)(b)), we find

$$
S_2^h = -\frac{1}{2}\sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{\sigma_{M,M'}^K\subset\partial M} G\left(u_M^{n+1}, u_{M'}^{n+1};\,\mathrm{d}V_{M,M'}^{n+1}\right)\left(\psi_{M'}^{n+1} - \psi_M^{n+1}\right)\left|\sigma_{M,M'}^K\right|.
$$

For each triplet of neighbors M, M', and M'' pick for $u_{K,\min}^{n+1}$ the quantity defined by

$$
u_{K,\min}^{n+1} = \min_{M\in\mathcal{M},\, M\cap K\neq\emptyset}\{u_M^{n+1}\}
$$

Set

$$
S_2^{h,*} := \sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{K,K\cap M\neq\emptyset} \sum_{M',M'\cap K\neq\emptyset} \chi\left(u_{K,\min}^{n+1}\right)\psi_M^{n+1}\,\mathrm{d}V_{M,M'}^{n+1}\left|\sigma_{M,M'}^K\right|.
$$

We have

$$
\begin{aligned}
S_2^{h,*} &= -\sum_{n=0}^{N} \Delta t_h \sum_{M\in\mathcal{M}_h} \sum_{K,K\cap M\neq\emptyset} \chi\left(u_{K,\min}^{n+1}\right)|K|\,\Lambda_K\psi_M^{n+1}\nabla\varphi_M\cdot\left(\sum_{M'',M''\cap K\neq\emptyset} v_{M''}^{n+1}\nabla\varphi_{M''}\right)\\
&= -\sum_{n=0}^{N} \Delta t_h \sum_{K\in\mathcal{T}_h} |K|\chi\left(u_{K,\min}^{n+1}\right)\Lambda_K\left(\sum_{M,M\cap K\neq\emptyset}\psi_M^{n+1}\nabla\varphi_M\right)\cdot\left(\sum_{M'',M''\cap K\neq\emptyset} v_{M''}^{n+1}\nabla\varphi_{M''}\right).
\end{aligned}
$$

Introduce $\overline{u}_h, \underline{u}_h$ defined by

$$\overline{u}_h|_{(t_n,t_{n+1}]\times K} := \max_{M,\, M\cap K\neq\emptyset}\{u_M^{n+1}\}, \qquad \underline{u}_h|_{(t_n,t_{n+1}]\times K} := \min_{M,\, M\cap K\neq\emptyset}\{u_M^{n+1}\}.$$

Consequently, we obtain

$$\begin{aligned} S_2^{h,*} &= -\sum_{n=0}^{N}\Delta t_h \sum_{K\in\mathcal{T}_h}\int_K \chi\left(\underline{u}_h\right)\Lambda_K\nabla v_{\mathcal{T}_h,\Delta t_h}^{n+1}\cdot(\nabla\psi)_{\mathcal{T}_h}^{n+1}\mathrm{d}\mathbf{x} \\ &= -\int_{Q_{t_\mathrm{f}}}\chi\left(\underline{u}_h\right)\Lambda\nabla v_{h,\Delta t_h}\cdot\nabla\psi_{\mathcal{T}_h,\Delta t_h}\mathrm{d}\mathbf{x}\,\mathrm{d}t. \end{aligned}$$

Next, we show that

$$\lim_{h\to 0}|S_2^h - S_2^{h,*}| = 0. \tag{3.44}$$

To do this, we begin by showing that $|\underline{u}_h - \overline{u}_h| \longrightarrow 0$ a.e. in Q_{t_f}.
By a slight adaptation of Lemma A.1 and thanks to estimate (3.33), one can conclude the existence of a constant $C>0$ (independent of h) such that

$$\begin{aligned} \int_0^{t_\mathrm{f}}\int_\Omega |A\left(\overline{u}_h\right) - A\left(\underline{u}_h\right)|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t &\le C\,h^2\int_0^{t_\mathrm{f}}\int_\Omega \left|\nabla A_{\mathcal{T}_h}\left(u_{\mathcal{T}_h,\Delta t_h}\right)(\mathbf{x},t)\right|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\le C\Lambda_-\,h^2\sum_{n=0}^{N}\Delta t_h\sum_{M\in\mathcal{M}_h}\sum_{\sigma_{M,M'}^K\subset\partial M}\Lambda_{M,M'}^K\left|A(u_{M'}^{n+1}) - A(u_M^{n+1})\right|^2 \le C\,h^2. \end{aligned}$$

Since A^{-1} is continuous, up to extraction of another subsequence, we deduce

$$|\overline{u}_h - \underline{u}_h| \to 0 \ \text{ a.e. in } Q_{t_\mathrm{f}}. \tag{3.45}$$

In addition, $\underline{u}_h \le u_{\mathcal{M}_h,\Delta t_h} \le \overline{u}_h$; moreover, by Lemma 3.12, $u_{\mathcal{M}_h,\Delta t_h}\to u$ a.e. in Q_{t_f}. Thus we see that $\chi(\underline{u}_h)\to\chi(u)$ a.e. in Q_{t_f} and in $L^p(Q_{t_\mathrm{f}})$, for $p<+\infty$. Using Lemma 3.12 again, the strong convergence of $\nabla\psi_{\mathcal{T}_h,\Delta t_h}$ towards $\nabla\psi$, and the weak convergence in $L^2\left(Q_{t_\mathrm{f}}\right)$ of $\nabla v_{\mathcal{T},\Delta t_h}$ towards ∇v, we conclude that

$$\lim_{h\to 0} S_2^{h,*} = -\int_0^{t_\mathrm{f}}\int_\Omega \chi(u)\Lambda\nabla v\cdot\nabla\psi\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

To prove (3.44), we remark that

$$\begin{aligned} &\left|G\left(u_M^{n+1},u_{M'}^{n+1};\,\mathrm{d}V_{M,M'}^{n+1}\right) - \chi\left(u_{K,\min}^{n+1}\right)\mathrm{d}V_{M,M'}^{n+1}\right| \\ &\qquad = \left|G\left(u_M^{n+1},u_{M'}^{n+1};\,\mathrm{d}V_{M,M'}^{n+1}\right) - G\left(u_{K,\min}^{n+1},u_{K,\min}^{n+1};\,\mathrm{d}V_{M,M'}^{n+1}\right)\right| \\ &\qquad \le \left|\mathrm{d}V_{M,M'}^{n+1}\right|\omega\left(2\left|u_{M'}^{n+1} - u_M^{n+1}\right|\right). \end{aligned}$$

Consequently, we obtain

$$\left|S_2^h - S_2^{h,*}\right| \le \int_0^{t_\mathrm{f}}\int_\Omega \omega\left(2\left|\overline{u}_h - \underline{u}_h\right|\right)\,\left|\nabla v_{\mathcal{T}_h,\Delta t_h}\cdot\nabla\psi_{\mathcal{T}_h,\Delta t_h}\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Applying the Cauchy-Schwarz inequality, and the convergence (3.44), we establish (3.43).
Finally, we note that it is easy to see that $S_4^h \xrightarrow[h,\Delta t_h\to 0]{} \displaystyle\int_{Q_{t_\mathrm{f}}} f\left(u\right)\psi\mathrm{d}\mathbf{x}\,\mathrm{d}t.$

3.7 Numerical simulation in two-dimensional space

In this section, we exhibit various two-dimensional numerical results provided by scheme (3.23)–(3.25) for the capture of spatial patterns for model (3.1) discussed in Section 3.2. Newton's algorithm is used to approach the solution U^{n+1} of the nonlinear system defined by equation (3.24), this algorithm is coupled with a bigradient method to solve the linear systems arising from the Newton algorithm as well as the linear system given by equation (3.25). Unless stated otherwise, throughout this section, we consider that the cell density is initially set as a spatially small random perturbation around the homogeneous steady state, and we assume zero-flux boundary conditions. The simulations are performed on an unstructured triangular mesh of the space domain $\Omega = (0, 10) \times (0, 10)$. We suppose that the species cells follow the logistic growth $f(u) = \mu u(1 - u/u_c)$, where μ is the *intrinsic growth rate*, and u_c is the *carrying capacity* of the population. Production term $g(u, v)$, squeezing probability $q(u)$, cell diffusivity $a(u)$, and chemotactic sensitivity $\chi(u)$ are given by (3.4), (3.7), and (3.8) respectively.

The pattern formation for model (3.1) with the associated functions mentioned above has been established in [78] using Turing's principle and the linear stability analysis, where the diffusion tensor is considered to be proportional to the identity matrix and the numerical simulations are carried on a one-dimensional space, while in [49], the same analysis is provided whereas the numerical simulations for the capture of spatial patterns are presented on a two-dimensional domain, and using the standard finite volume scheme.

The nontrivial uniform steady state of system (3.1) is given by $(u_s, v_s) = (u_c, \alpha u_c/\beta)$, and through the pattern formation analysis provided in [78, 49], the instability region of this steady state is determined by the following condition :

$$\mu + \beta a(u_c) - \alpha\chi(u_c) < -2\sqrt{\mu\beta a(u_c)}. \tag{3.46}$$

In order to verify the effectiveness of the proposed scheme, we consider three tests for the capture of spatio-temporal patterns for model (3.1) with different diffusion tensors. For each test, we choose a set of parameters, such that the instability condition (3.46) is satisfied. We fix $d_1 = 0.25$, $u_c = 0.25$, $\overline{u} = 1.0$, $\mu = 0.5$, $\alpha = 10.0$, $\beta = 10.0$, $\zeta = 20$, and $\gamma = 3$. On the other hand, and in the definition of the numerical flux function G defined by (3.20), we take

$$\chi_{\uparrow}(z) = \chi(\min\{z, \tilde{u}\}) \text{ and } \chi_{\downarrow}(z) = \chi(\max\{z, \tilde{u}\}) - \chi(\tilde{u}),$$

where $\tilde{u} = \frac{\overline{u}}{\sqrt[\gamma]{\gamma+1}}$. Finally, we consider a small time step $\Delta t = 0.005$ and a nonuniform primary mesh with small refinement consisting of 14 336 triangles. Thus, the associated Donald dual mesh consists of 7297 dual control volumes.

Test 1 (Isotropic case) Here, we establish the generation of spatial patterns for the volume-filling chemotaxis model (3.1) with homogeneous diffusion tensors. In this test, we take $\Lambda(\mathbf{x}) = D(\mathbf{x}) = I_2(\mathbf{x})$ and initially the chemical concentration is set to be a constant equal to v_s.

Figures 3.2–3.3 show for different moments, the pattern formation for model (3.1) with identity diffusion tensors. We see that the random distribution of the cell density leads to a merging process in all directions of the space at $t = 0.55$s, which continues for $t = 2.35$s then it stops when the time $t \geq 20.75$s, and new stationary spot patterns appear as shown at $t = 40$s.

Time evolution of the cell density. Here we consider the time evolution of the cell density at fixed points in the right snapshot of Figure 3.3. Indeed, we want to show that the cell density stabilizes at a certain moment ; hence, we prove that the volume-filling chemotaxis model (3.1) generates stationary spatial patterns. Figure 3.4 shows the evolution of the cell density with respect

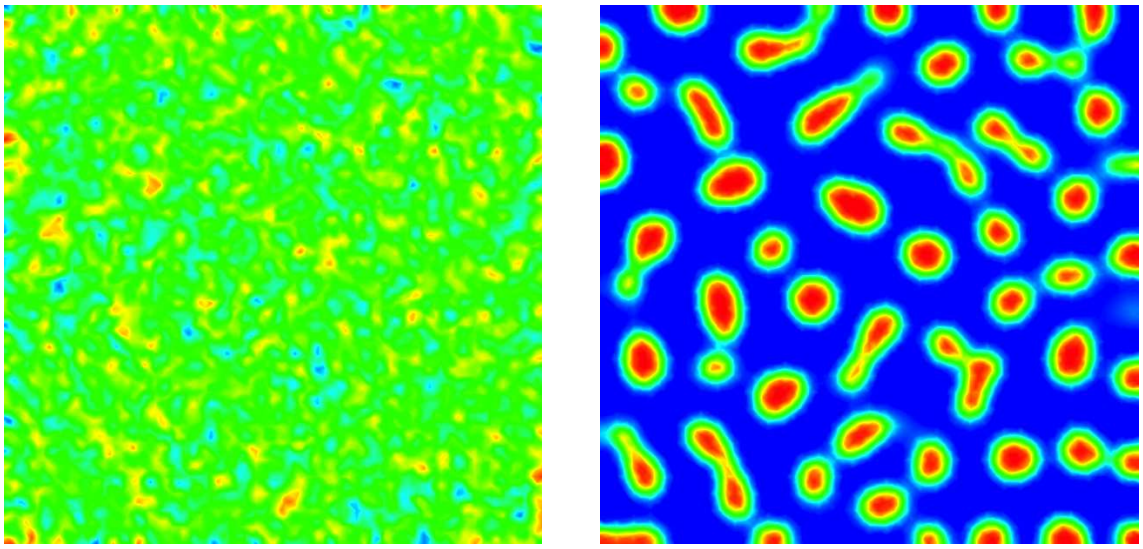


FIGURE 3.2 – Pattern formation for the full chemotaxis model (3.1) on a 2-D domain $\Omega = (0, 10)\times(0, 10)$ at time $t = 0\,\text{s}$ with $0 \leq u \leq 1$ (*left*) and at time $t = 0.55\,\text{s}$ with $3.10^{-3} \leq u \leq 0.99$ (*right*).

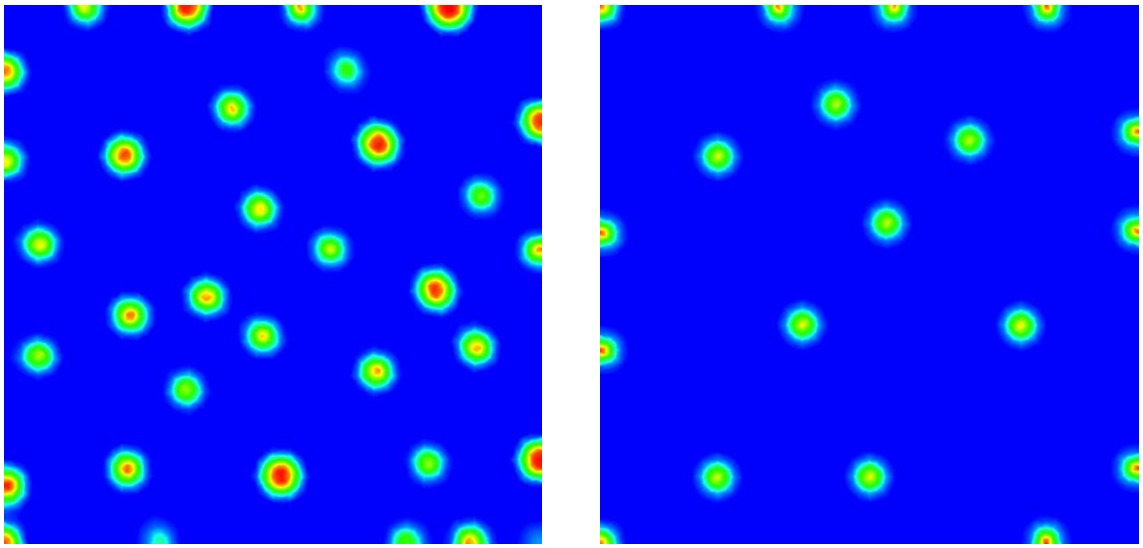


FIGURE 3.3 – Pattern formation for the full chemotaxis model (3.1) at time $t = 2.35\,\text{s}$ with $1.9\times10^{-4} \leq u \leq 0.98$ (*left*) and at time $t = 40\,\text{s}$ with $3.9\times10^{-3} \leq u \leq 0.832$ (*right*). The red dots (or rods) represent the cell aggregation where cell density is higher than that of the blue area.

to the time at point $P_1\,(5.25; 6)$ in the red line, at point $P_2\,(5; 1.25)$ in the green line, and at point $P_3\,(4.35; 8.1)$ in the blue line. We observe that the cell density at these points increases and then decreases with response to the gradient of the chemoattractant which plays an essential role to stop the aggregation of the cells. Next, the cell density stabilizes for all points when t is greater or equal to $13\,\text{s}$. We note that the same results are obtained for the other spot patterns ; however, for the sake of brevity, they are not provided here.

Test 2 (Anisotropic case) In this test, we take a homogeneous anisotropic diffusion tensors of the form $\begin{bmatrix} 1 & 0 \\ 0 & \xi \end{bmatrix}$. We investigate the pattern formation for model (3.1) and consider that the

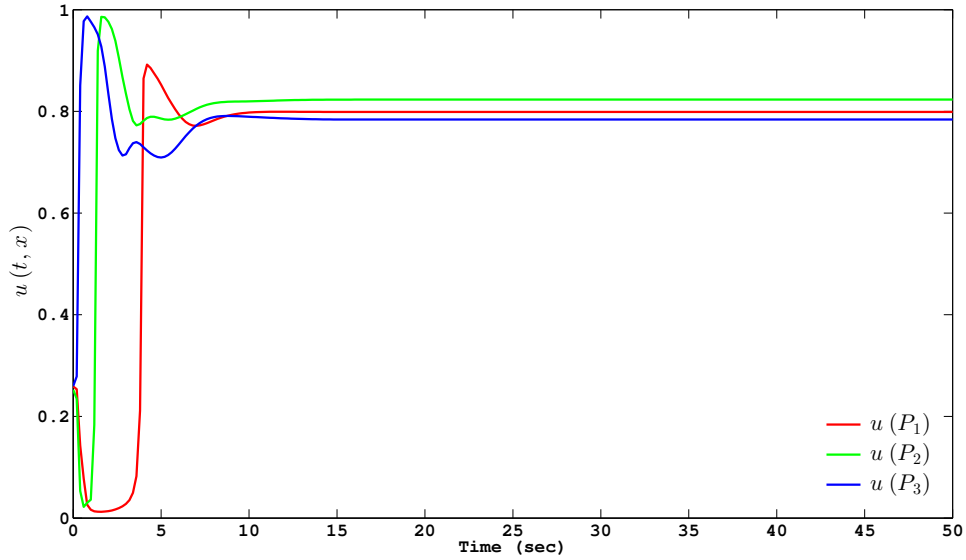


FIGURE 3.4 – Evolution of the cell density with respect to the time at three different points : $P_1\,(5.25; 6)$ in the red line, $P_2\,(5; 1.25)$ in the green line, $P_3\,(4.35; 8.1)$ in the blue line.

particles diffuse more rapidly in the $\mathbf{x}$-axis direction than the $\mathbf{y}$-axis direction. In other words, we are concerned by a matrix with high diffusivity in the horizontal direction and low diffusivity in the orthogonal direction ; for that we take $\xi < 1$, say for example $\xi = 0.5$. We pick up snapshots for the pattern formation at the same moments as those of the isotropic case.

Figures 3.5–3.6 show the evolution of spatial patterns for model (3.1) for different time moments. We observe that we have the same results as before (same patterning, and emerging process), except that more spot patterns are obtained and they are stretched in the horizontal direction as shown in the last snapshot in figure 3.6.

The time evolution for the last plot in figure 3.6 for different spots is given in figure 3.7. It shows that the stationary spot patterns are obtained when $t \geq 23.5$s.

Test 3 (Heterogeneous anisotropic case) In this test, we decompose the domain Ω into two regions Ω_1 and Ω_2, where $\Omega_1 = (0, 10) \times (0, 5]$ and $\Omega_2 = (0, 10) \times (5, 10)$. Moreover, we assume that the diffusion tensors are anisotropic and heterogeneous, and are given by :

$$\Lambda(\mathbf{x}) = D(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & \lambda(\mathbf{x}) \end{pmatrix}, \text{ with } \begin{cases} \lambda(\mathbf{x}) = 0.5, & \text{if } \mathbf{x} \in \Omega_1, \\ \lambda(\mathbf{x}) = 1.5, & \text{if } \mathbf{x} \in \Omega_2. \end{cases}$$

Figures 3.8–3.9 show the evolution of spatial patterns for model (3.1) for different time moments.
In region Ω_2, where diffusion is more interesting in the $\mathbf{y}$-axis direction, and at $t = 0.5$ s, we see that the aggregations form quickly, and they are much larger than those formed in region Ω_1, which means that high diffusion in region Ω_2 (compared to that in region Ω_1) accelerate the merging process. On the other hand, at $t = 49$ s the aggregations in region Ω_2 disappear to form 5 spatial patterns ; whereas, a plenitude of spatial patterns is seen in region Ω_1 compared to region Ω_2. Therefore, there exists a high dependence between the rate of diffusion and the generation of spatial patterns.

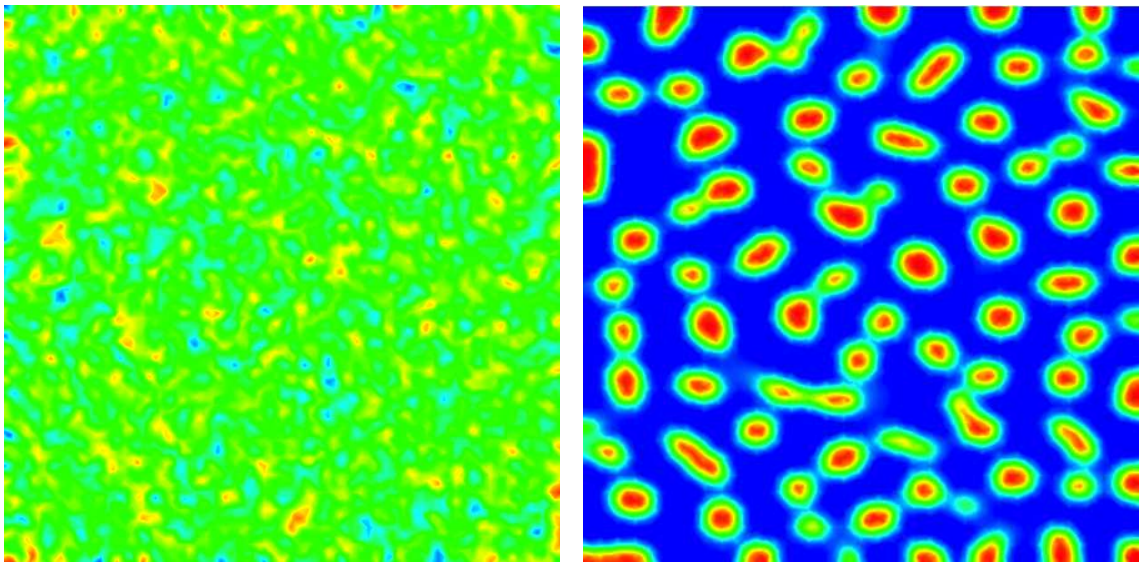


FIGURE 3.5 – Pattern formation for the full chemotaxis model (3.1) on a 2-D domain $\Omega = (0,10)\times(0,10)$ at time $t = 0\,\mathrm{s}$ with $0 \leq u \leq 1$ (*left*) and at time $t = 0.55\,\mathrm{s}$ with $4.10^{-3} \leq u \leq 0.98$ (*right*).

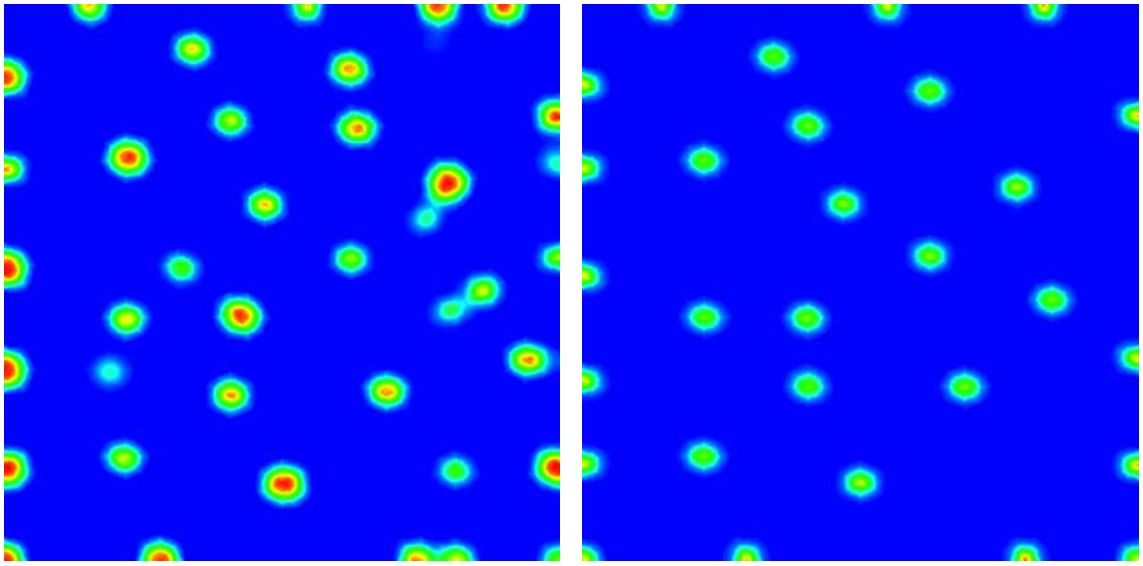


FIGURE 3.6 – Pattern formation for the full chemotaxis model (3.1) at time $t = 2.35\,\mathrm{s}$ with $3.2\times10^{-4} \leq u \leq 0.98$ (*left*) and at time $t = 40\mathrm{s}$ with $4\times10^{-3} \leq u \leq 0.827$ (*right*).

The time evolution for the last plot of figure 3.9 for different spots is given in figure 3.10. It shows that stationary spot patterns are obtained when $t \geq 125\mathrm{s}$.

Comparing figures 3.5-3.9, one can deduce that the patterning depends on diffusion coefficients, it is also known to depend on the size of the domain. These results prove the robustness of the *control volume finite element scheme* to capture spatial patterns for a volume-filling chemotaxis model with anisotropic diffusion tensors.

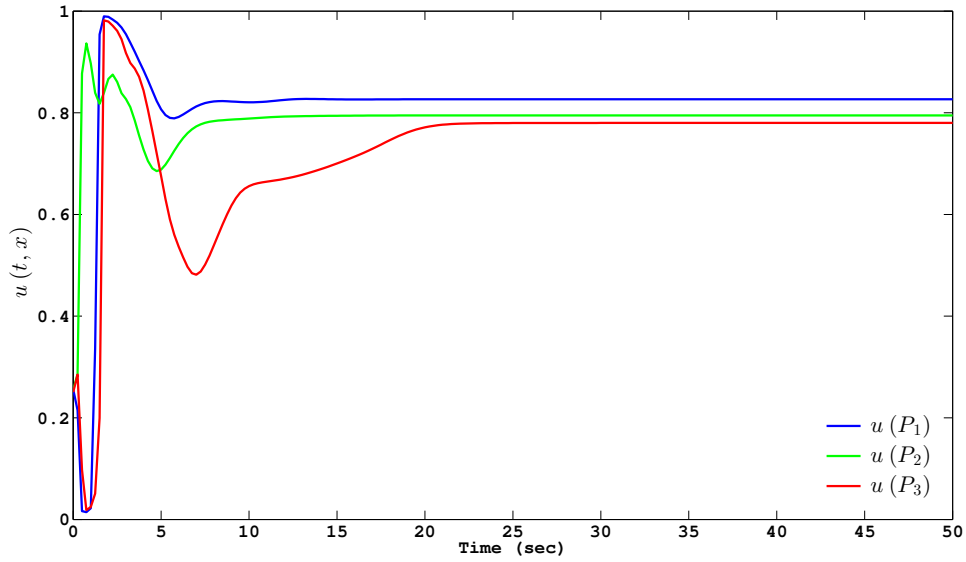


FIGURE 3.7 – The evolution of the cell density with respect to the time at three different points : $P_1\,(5;1.4)$ in the blue line, $P_2\,(6.25;5.45)$ in the green line, $P_3\,(9.9;8.05)$ in the red line.

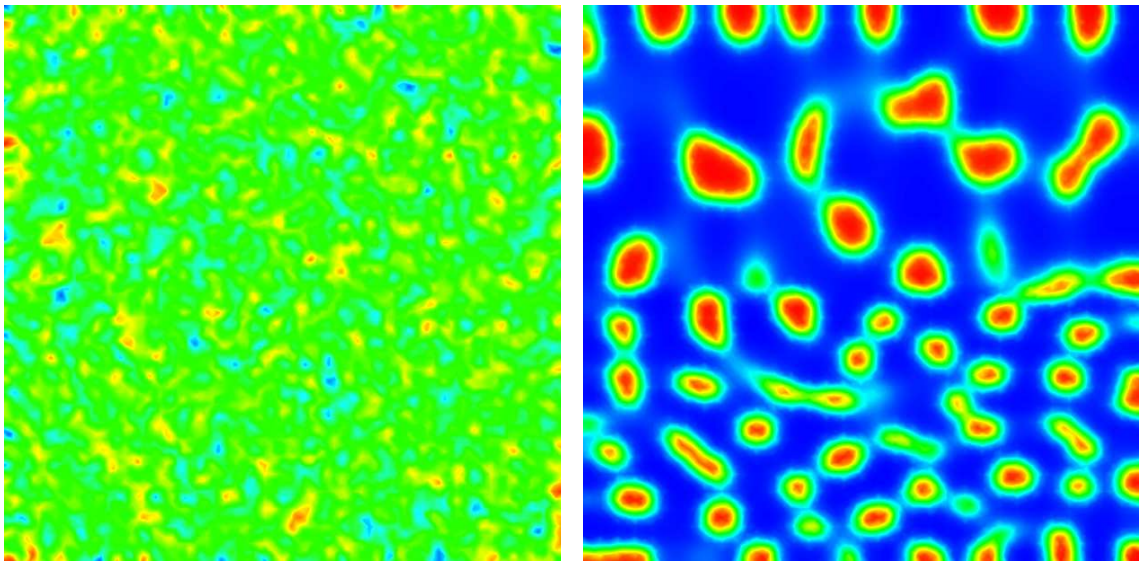


FIGURE 3.8 – Pattern formation for the full chemotaxis model (3.1) on a 2-D domain $\Omega = (0,10)\times(0,10)$ at time $t = 0\,\mathrm{s}$ with $0 \leq u \leq 1$ (*left*) and at time $t = 0.5\,\mathrm{s}$ with $6.93 \times 10^{-3} \leq u \leq 0.99$ (*right*).

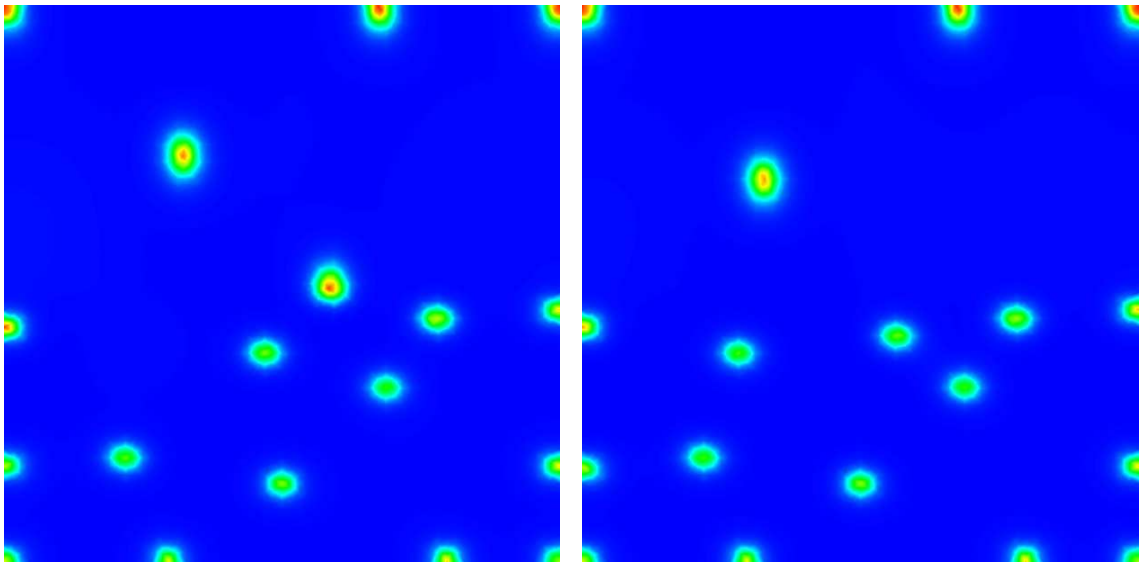


FIGURE 3.9 – Pattern formation for the full chemotaxis model (3.1) at time $t = 49\text{s}$ with $3.67\times 10^{-3} \leq u \leq 0.88$ (*left*) and at time $t = 150\text{s}$ with $4.3\times 10^{-3} \leq u \leq 0.8447$ (*right*).

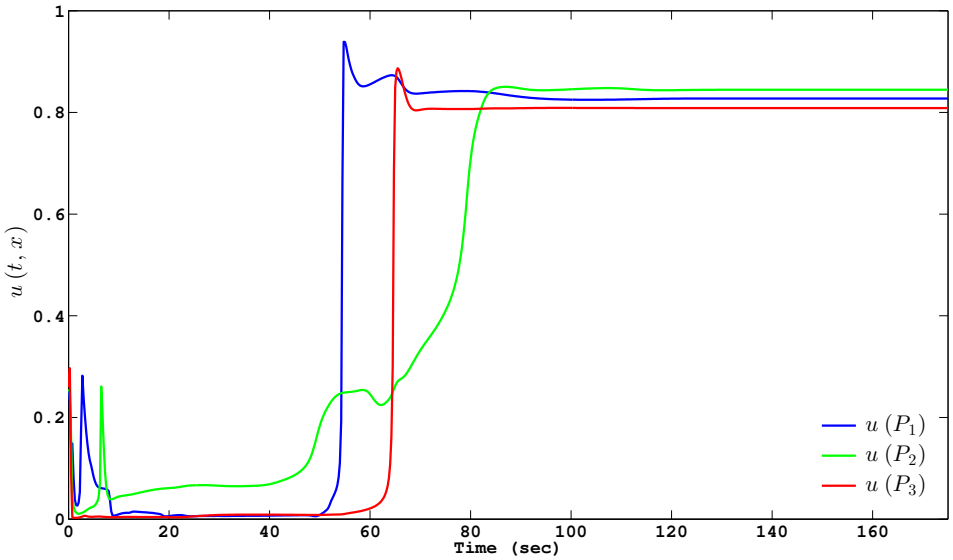


FIGURE 3.10 – The evolution of the cell density with time at three different points : $P_1\,(5.70; 4.05)$ in the blue line, $P_2\,(3.28; 6.88)$ in the green line, $P_3\,(2.80; 3.75)$ in the red line.

4

A nonlinear CVFE scheme for the modified degenerate anisotropic chemotaxis model

Sommaire

4.1 Introduction

In this chapter, we are interested in a degenerate nonlinear parabolic reaction–convection–diffusion system modeling the chemotaxis process over general triangular mesh, with anisotropic and heterogeneous diffusion tensors.

In chapter 3, the convergence analysis of the mixed conforming piecewise linear finite elements on triangles for the diffusion term and finite volume on dual elements for the other terms has carried out for the case of anisotropic and heterogeneous diffusion problems under an essential assumption that all the transmissibility coefficients are nonnegative. However, there is no sufficient conditions for nonnegativity of transmissibility coefficients and therefore the schemes do not permit to tackle general anisotropic diffusion problems.

Recently, Cancès and Guichard proposed and analyzed in [16] a nonlinear *Control Volume Finite Element* (CVFE) scheme for solving degenerate anisotropic parabolic diffusion equations modeling flows in porous media. The convergence analysis is carried out without any restriction on the transmissibility coefficients, and the effectiveness of the scheme is tested using anisotropic diffusion tensors over an unstructured mesh.

The purpose of this chapter is to extend the idea of [16] to a fully nonlinear degenerate parabolic reaction–convection–diffusion system modeling the chemotaxis process over general mesh, with anisotropic and heterogeneous diffusion tensors. Here, the applied discretization is very similar to that one employed in [50] ; the diffusion terms are discretized by means of a conforming piecewise linear finite element method on a primal triangular mesh and using the Godunov scheme to approximate the diffusion fluxes provided by the conforming finite element reconstruction. The others terms are discretized by means of a nonclassical upwind finite volume method on a dual mesh (Donald mesh or Median dual mesh).

The rest of this chapter is organized as follows. In Section 4.2, we introduce the modified Keller-Segel model as well as the assumptions made about the system. In Section 4.3 and in order to discretize the domain, we define a primal triangular mesh and its corresponding dual barycentric mesh, we then define two different reconstructions of the discrete solutions based on the spatial discretizations ; the first stands to the usual $\mathbb{P}_1$ finite element reconstruction while the second stands to the piecewise constant reconstruction. In Section 4.4, we define the discretization of the diffusion and convection terms and introduce the nonlinear CVFE scheme. With the help of some technical lemmas in Section 4.5, we prove the existence of a discrete solution to the CVFE scheme based on the establishment of *a priori* estimates on the discrete solution as well as the discrete maximum principle. In Section 4.6, we give estimates on differences of time and space translates for the approximate solutions. In Section 4.7, using the Kolmogorov relative compactness criterion, we prove the convergence of a subsequent of the sequence of discrete solutions to a weak solution of the continuous problem. Finally, some numerical simulations are carried out, in Section 4.8, to show the effectiveness of the scheme to tackle degenerate anisotropic and heterogeneous diffusion problems over general unstructured mesh.

4.2 The modified chemotaxis model

Let Ω be an open bounded connected polygonal domain of $\mathbb{R}^2$, and $t_{\mathrm{f}} > 0$ be a fixed time. The modified Keller-Segel system modeling the chemotaxis process is given by the following set of equations

$$\begin{cases} \partial_t u - \operatorname{div}\left(\Lambda(\mathbf{x})\, a(u)\, \nabla u - \Lambda(\mathbf{x})\, \chi(u)\, \nabla v\right) = f(u) & \text{in } Q_{t_{\mathrm{f}}} = \Omega \times (0, t_{\mathrm{f}}), \\ \partial_t v - \operatorname{div}\left(D(\mathbf{x})\, \nabla v\right) = g(u, v) & \text{in } Q_{t_{\mathrm{f}}} = \Omega \times (0, t_{\mathrm{f}}). \end{cases} \tag{4.1}$$

The system is complemented with homogeneous zeros-flux boundary conditions on $\Sigma_{t_f} := \partial\Omega \times (0, t_f)$ given by

$$(\Lambda(\mathbf{x})\, a(u)\, \nabla u - \Lambda(\mathbf{x})\, \chi(u)\, \nabla v) \cdot \mathbf{n} = 0, \qquad D(\mathbf{x})\, \nabla v \cdot \mathbf{n} = 0, \tag{4.2}$$

and the initial conditions on Ω :

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad v(\mathbf{x}, 0) = v_0(\mathbf{x}). \tag{4.3}$$

In the above model, the density of the cell-population and the chemoattractant concentration are represented by $u = u(\mathbf{x}, t)$ and $v = v(\mathbf{x}, t)$ respectively. Next, $a(u)$ is a density-dependent diffusion coefficient, and $\Lambda(\mathbf{x})$ is the diffusion tensor in a heterogeneous medium. Furthermore, the function $\chi(u)$ is the chemoattractant sensitivity, and $D(\mathbf{x})$ is the diffusion tensor for v. The function $f(u)$ describes cell proliferation and cell death. The function $g(u, v)$ describes the rates of production and degradation of the chemoattractant ; here, we assume it is the linear function given by

$$g(u, v) = \alpha u - \beta v, \qquad \alpha, \beta \geq 0. \tag{4.4}$$

The unit normal vector to $\partial\Omega$ outward to Ω is denoted by $\mathbf{n}$.
Let us state the main assumptions made about model (4.1)–(4.3) :

(A1) The cell-density diffusion $a : [0, 1] \longrightarrow \mathbb{R}^+$ is a continuous function such that, $a(0) = a(1) = 0$, and $a(u) > 0$ for $0 < u < 1$.

(A2) The chemosensitivity $\chi : [0, 1] \longrightarrow \mathbb{R}^+$ is a continuous function such that, $\chi(0) = \chi(1) = 0$. Furthermore, we assume that there exists a function $\mu \in C([0, 1]; \mathbb{R}^+)$, such that $\mu(u) = \dfrac{\chi(u)}{a(u)}$ and $\mu(0) = \mu(1) = 0$.

(A3) The diffusion tensors Λ and D are two bounded, uniformly positive symmetric tensors on Ω, that is : $\forall \mathbf{w} \neq 0$, there exists two nonnegative constants T_- and T_+ such that $0 < T_- |\mathbf{w}|^2 \leq \langle T(\mathbf{x})\mathbf{w}, \mathbf{w} \rangle \leq T_+ |\mathbf{w}|^2 < \infty, \quad T = \Lambda$ or D.

(A4) The cell density proliferation f is a continuous function such that $f(0) \geq 0$ and $f(1) \leq 0$.

(A5) The initial function u_0 and v_0 are two functions in $L^2(\Omega)$ such that, $0 \leq u_0 \leq 1$ and $v_0 \geq 0$.

In the sequel, we use the Lipschitz continuous nondecreasing function $\xi : [0, 1] \longrightarrow \mathbb{R}$ defined by

$$\xi(u) := \int_0^u \sqrt{a(s)}\, \mathrm{d}s, \qquad \forall u \in \mathbb{R}. \tag{4.5}$$

We recall the definition of a weak solution of system (4.1)–(4.3).

Definition 4.1 (weak solution)**.** Under the assumptions (A1)–(A5), we say that the couple of measurable functions (u, v) is a weak solution of the system (4.1)–(4.3) if

$$\begin{aligned} &0 \leq u(\mathbf{x}, t) \leq 1, \qquad 0 \leq v(\mathbf{x}, t) \quad \text{for a.e. in } Q_{t_f}, \\ &\xi(u) \in L^2\left(0, t_f; H^1(\Omega)\right), \\ &v \in L^\infty(Q_{t_f}) \cap L^2\left(0, t_f; H^1(\Omega)\right), \end{aligned}$$

and for all $\varphi, \psi \in \mathcal{D}\left(\overline{\Omega} \times [0, t_{\mathrm{f}})\right)$, one has

$$
\begin{aligned}
-\int_{\Omega} u_0(\mathbf{x})\,\varphi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} - \iint_{Q_{t_{\mathrm{f}}}} u\,\partial_t\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \iint_{Q_{t_{\mathrm{f}}}} \sqrt{a(u)}\Lambda(\mathbf{x})\,\nabla\xi(u)\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\
- \iint_{Q_{t_{\mathrm{f}}}} \Lambda(\mathbf{x})\chi(u)\,\nabla v\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = \iint_{Q_{t_{\mathrm{f}}}} f(u)\,\varphi(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \\
-\int_{\Omega} v_0(\mathbf{x})\,\psi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} - \iint_{Q_{t_{\mathrm{f}}}} v\,\partial_t\psi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \iint_{Q_{t_{\mathrm{f}}}} D(\mathbf{x})\,\nabla v\cdot\nabla\psi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\
= \iint_{Q_{t_{\mathrm{f}}}} g(u,v)\,\psi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.
\end{aligned}
$$

4.3 Space-time discretization and notations

In this section, we describe the space and time discretizations of $Q_{t_{\mathrm{f}}}$, define the approximation spaces, introduce useful properties on discrete H^1-norms stemming from finite elements discretizations.

4.3.1 Space discretizations of Ω.

In order to discretize problem (4.1)–(4.2), we perform a finite element triangulation $\mathcal{T}$ of the polygonal domain Ω, consisting of open bounded triangles such that $\overline{\Omega} = \bigcup_{T\in\mathcal{T}} \overline{T}$ and such that for all $T, T' \in \mathcal{T}$, $\overline{T} \cap \overline{T'}$ is either an empty set or a common vertex or edge of T and T'. We denote by $\mathcal{V}$ the set of vertices of the discretization $\mathcal{T}$, located at positions $(\mathbf{x}_K)_{K\in\mathcal{V}}$, and by $\mathcal{E}$ the set of edges of $\mathcal{T}$ joining two vertices of $\mathcal{V}$, that are contained in hyperplanes of $\mathbb{R}^d$. The edge joining two vertices K and L is denoted by σ_{KL}.

For a given triangle $T \in \mathcal{T}$, we denote by $\mathbf{x}_T$ the centre of gravity of T, by $\mathcal{E}_T$ the set of the edges of T, by $h_T = \mathrm{diam}(T)$ the diameter of T, and by ρ_T the diameter of the largest ball inscribed in the triangle T. We denote by h the size of the triangulation $\mathcal{T}$ and by $\theta_{\mathcal{T}}$ the shape regularity assumption on the triangulation $\mathcal{T}$, they are given by

$$
h_{\mathcal{T}} := \max_{T\in\mathcal{T}} h_T, \quad \theta_{\mathcal{T}} := \max_{T\in\mathcal{T}} \frac{h_T}{\rho_T}.
$$

For $K \in \mathcal{V}$, we denote by $\mathcal{E}_K$ the set of the edges having K as an extremity, and by $\mathcal{T}_K$ the subset of $\mathcal{T}$ including the triangles having K as a vertex. We also define a dual barycentric mesh $\mathcal{M}$ (known as Donald dual or Median dual mesh) generated by the triangulation mesh $\mathcal{T}$. There is one dual element ω_K associated with each vertex $K \in \mathcal{V}$. We construct it around the vertex K by connecting the barycenter $\mathbf{x}_T$ of each triangle $T \in \mathcal{T}_K$ with the barycenters $\mathbf{x}_\sigma$ of the edges $\sigma \in \mathcal{E}_K$. We refer to Fig. 4.1 for an illustration of the primary and the dual barycentric mesh in a two-dimensional space. Note that $\overline{\Omega} = \bigcup_{K\in\mathcal{V}} \overline{\omega}_K$. The 2-dimensional Lebesgue measure of ω_K is denoted by m_K.

4.3.2 Discrete finite elements space $\mathcal{H}_{\mathcal{T}}$, control volumes space $\mathcal{X}_{\mathcal{M}}$.

We define two discrete functional spaces associated with each mesh of the above constructed meshes. The first one, denoted by $\mathcal{H}_{\mathcal{T}}$, is the usual $\mathbb{P}_1$-finite element space corresponding to the triangular mesh $\mathcal{T}$, consisting of piecewise affine finite elements.

$$
\mathcal{H}_{\mathcal{T}} := \left\{\varphi \in C^0\left(\overline{\Omega}\right);\, \varphi_{|_T} \in \mathbb{P}_1(\mathbb{R}),\ \forall T \in \mathcal{T}\right\} \subset H^1(\Omega).
$$

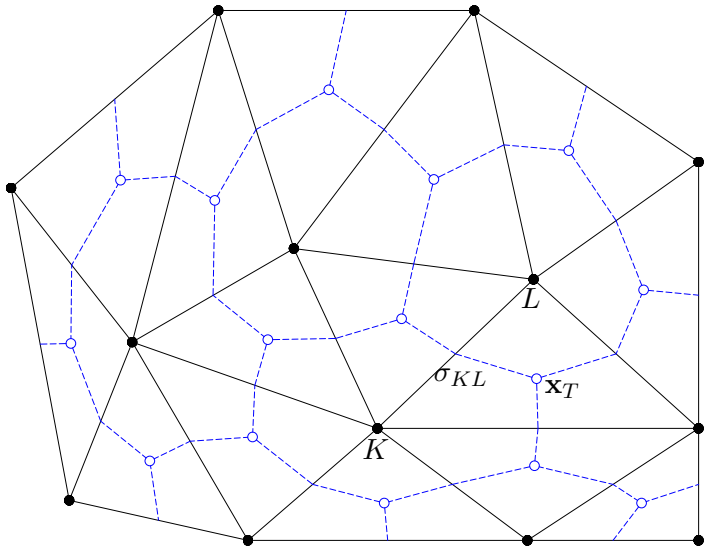


FIGURE 4.1 – Triangular mesh $\mathcal{T}$ and Donald dual mesh $\mathcal{M}$: dual volumes, vertices, interfaces.

The canonical basis of $\mathcal{H}_\mathcal{T}$ is spanned by the shape functions $(\varphi_K)_{K\in\mathcal{V}}$, such that

$$\varphi_K(\mathbf{x}_K) = 1, \quad \varphi_K(\mathbf{x}_L) = 0, \quad \text{if } L \neq K, \quad \forall K \in \mathcal{V}.$$

On the other hand, we denote by $\mathcal{X}_\mathcal{M}$ the discrete control volumes space consisting of piecewise constant functions on the dual mesh $\mathcal{M}$.

$$\mathcal{X}_\mathcal{M} = \{\varphi : \Omega \longrightarrow \overline{\mathbb{R}} \text{ measurable} ; \varphi_{|\omega_K} \in \overline{\mathbb{R}} \text{ is constant}, \quad \forall K \in \mathcal{V}\}.$$

Given a vector $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$(resp. $(v_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$), there exists a unique function $u_\mathcal{T} \in \mathcal{H}_\mathcal{T}$ (resp. $v_\mathcal{T} \in \mathcal{H}_\mathcal{T}$) and a unique $u_\mathcal{M} \in \mathcal{X}_\mathcal{M}$ (resp. $v_\mathcal{M} \in \mathcal{X}_\mathcal{M}$) such that

$$\begin{aligned} u_\mathcal{T}(\mathbf{x}_K) = u_\mathcal{M}(\mathbf{x}_K) = u_K, \qquad \forall K \in \mathcal{V}, \\ v_\mathcal{T}(\mathbf{x}_K) = v_\mathcal{M}(\mathbf{x}_K) = v_K, \qquad \forall K \in \mathcal{V}. \end{aligned} \tag{4.6}$$

4.3.3 Time discretization of $(0, t_\mathrm{f})$.

For the time discretization of the interval $(0, t_\mathrm{f})$, we do not impose any restriction on the time step, and we restrict our study to the case of a uniform time discretization. In addition, we assume that the spatial meshes do not change with the time step. We note that all the results presented in this chapter can be extended to the case of general time discretization.
Let N be a nonnegative integer, we define the uniform time step $\Delta t = t_\mathrm{f}/(N+1)$, and $t_n = n\Delta t$ for all $n \in \{0, \cdots, N+1\}$, so that $t_0 = 0$, and $t_{N+1} = t_\mathrm{f}$.

4.3.4 Space-time discretization of Q_{t_f}.

Here, we define the space and time discrete spaces $\mathcal{H}_{\mathcal{T},\Delta t}$ and $\mathcal{X}_{\mathcal{M},\Delta t}$ as the set of piecewise constant functions in time with values in $\mathcal{H}_\mathcal{T}$ and $\mathcal{X}_\mathcal{T}$ respectively.

$$\begin{aligned} \mathcal{H}_{\mathcal{T},\Delta t} &= \{\varphi \in L^2\left(0, t_\mathrm{f}; H^1(\Omega)\right), \varphi(\mathbf{x}, t) = \varphi(\mathbf{x}, t_{n+1}) \in \mathcal{H}_\mathcal{T}, \quad \forall t \in (t_n, t_{n+1}]\}, \\ \mathcal{X}_{\mathcal{M},\Delta t} &= \{\varphi : Q_{t_\mathrm{f}} \longrightarrow \overline{\mathbb{R}} \text{ measurable}, \varphi(\mathbf{x}, t) = \varphi(\mathbf{x}, t_{n+1}) \in \mathcal{X}_\mathcal{M}, \quad \forall t \in (t_n, t_{n+1}]\}. \end{aligned}$$

For a given $(u_K^n)_{n\in\{0,\cdots,N+1\},K\in\mathcal{V}} \in \mathbb{R}^{(N+2)\#\mathcal{V}}$ (resp. $(v_K^n)_{n\in\{0,\cdots,N+1\},K\in\mathcal{V}}$), there exists a unique function $u_{\mathcal{T},\Delta t} \in \mathcal{H}_{\mathcal{T},\Delta t}$ (resp. $v_{\mathcal{T},\Delta t} \in \mathcal{H}_{\mathcal{T},\Delta t}$) and a unique $u_{\mathcal{M},\Delta t} \in \mathcal{X}_{\mathcal{M},\Delta t}$ (resp. $v_{\mathcal{M},\Delta t} \in \mathcal{X}_{\mathcal{M},\Delta t}$) such that

$$\begin{aligned} u_{\mathcal{T},\Delta t}(\mathbf{x}_K,t) = u_{\mathcal{M},\Delta t}(\mathbf{x}_K,t) = u_K^{n+1}, \qquad \forall K\in\mathcal{V},\ \forall t\in(t_n,t_{n+1}],\\ v_{\mathcal{T},\Delta t}(\mathbf{x}_K,t) = v_{\mathcal{M},\Delta t}(\mathbf{x}_K,t) = v_K^{n+1}, \qquad \forall K\in\mathcal{V},\ \forall t\in(t_n,t_{n+1}]. \end{aligned} \tag{4.7}$$

4.3.5 Main property

For all $(K,L)\in\mathcal{V}^2$, we define the coefficient $\mathbf{T}_{KL}$ by

$$\mathbf{T}_{KL} = -\int_\Omega \mathbf{T}(\mathbf{x})\nabla\varphi_K(\mathbf{x})\cdot\nabla\varphi_L(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathbf{T}_{LK}, \qquad \mathbf{T}(\mathbf{x}) = \Lambda(\mathbf{x}) \text{ or } D(\mathbf{x}). \tag{4.8}$$

We have $\mathbf{T}_{KK} = -\sum_{L\neq K}\mathbf{T}_{KL}$, since $\sum_{K\in\mathcal{V}}\nabla\varphi_K = 0$. As a consequence, given $u_\mathcal{T}$ and $v_\mathcal{T}$ two elements of $\mathcal{H}_\mathcal{T}$, one has

$$\int_\Omega \mathbf{T}(\mathbf{x})\nabla u_\mathcal{T}\cdot\nabla v_\mathcal{T}\mathrm{d}\mathbf{x} = \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(u_K-u_L)(v_K-v_L), \quad \mathbf{T}(\mathbf{x}) = \Lambda(\mathbf{x}) \text{ or } D(\mathbf{x}).$$

Proof. According to the definition of the approximate functions $u_\mathcal{T}$ and $v_\mathcal{T}$, one gets

$$\begin{aligned} \int_\Omega \mathbf{T}(\mathbf{x})\nabla u_\mathcal{T}\cdot\nabla v_\mathcal{T}\mathrm{d}\mathbf{x} &= \sum_{T\in\mathcal{T}}\int_T \mathbf{T}(\mathbf{x})\nabla u_\mathcal{T}\cdot\nabla v_\mathcal{T}\mathrm{d}\mathbf{x}\\ &= \sum_{T\in\mathcal{T}} |T|\,\mathbf{T}_T\left(\sum_{K\in\mathcal{V}} u_K\nabla\varphi_{K_{|T}}\right)\cdot\left(\sum_{L\in\mathcal{V}} v_L\nabla\varphi_{L_{|T}}\right)\\ &= -\sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}u_Kv_L - \sum_{K\in\mathcal{V}}\mathbf{T}_{KK}u_Kv_K\\ &= -\sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}u_Kv_L + \sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}u_Kv_K\\ &= \sum_{K\in\mathcal{V}}\sum_{L\in\mathcal{V},L\neq K}\mathbf{T}_{KL}(v_K-v_L)u_K\\ &= \sum_{\sigma_{KL}\in\mathcal{E}}\mathbf{T}_{KL}(v_K-v_L)(u_K-u_L), \end{aligned}$$

where, we have used the discrete integration by parts. □

4.4 The nonlinear CVFE scheme

In this section, we give the discretization of system (4.1)–(4.3). The discretization defined here, is very similar to that given in chapter 3. We consider an implicit Euler scheme in time for the time evolution terms. On the other hand, the approximations in space are obtained using either the finite element approach in compliance with the Godunov scheme, or a nonclassical upwind finite volume scheme.

4.4.1 Discretization of the first equation of system (4.1)

Let us begin by introducing the discretization of the diffusive term. To do that we consider the following diffusion equation

$$\partial_t u - \operatorname{div}\left(\Lambda(\mathbf{x})\, a(u)\, \nabla u\right) = 0. \tag{4.9}$$

By the change of variables $\vec{\mathbf{V}} = -\nabla u$, equation (4.9) becomes

$$\partial_t u + \operatorname{div}\left(\Lambda(\mathbf{x})\, a(u)\, \vec{\mathbf{V}}\right) = 0,$$

which has the form of an hyperbolic PDE equation. Therefore the Godunov scheme, which was introduced in [41], may be written at the interface $\sigma_{KL} \in \mathcal{E}$ by the following expression

$$F_{KL}(u_K, u_L) = \begin{cases} \min\limits_{u\in[u_K,u_L]} \Lambda_{KL}(u_K - u_L)\, a(u), & \text{if } u_K \le u_L, \\ \max\limits_{u\in[u_L,u_K]} \Lambda_{KL}(u_K - u_L)\, a(u), & \text{if } u_L \le u_K, \end{cases}$$

where $\Lambda_{KL}(u_K - u_L)$ represents the approximation of the flux $-\Lambda\nabla u\cdot\mathbf{n}_{KL}$ at the interface σ_{KL}. F_{KL} is called the numerical flux function, it verifies the monotonicity property (see e.g. [31]. Note that, for every real number $\kappa \in \mathbb{R}$ and for every real function f, one has

$$\min(\kappa f) = \begin{cases} \kappa \min f, & \text{if } \kappa \ge 0, \\ \kappa \max f, & \text{if } \kappa < 0, \end{cases} \qquad \max(\kappa f) = \begin{cases} \kappa \max f, & \text{if } \kappa \ge 0, \\ \kappa \min f, & \text{if } \kappa < 0. \end{cases}$$

As a consequence, one can conclude that the monotone flux scheme F_{KL} may be summarized by the following form

$$F_{KL}(u_K, u_L) = \begin{cases} \Lambda_{KL}(u_K - u_L) \max\limits_{u\in[u_K,u_L]} a(u), & \text{if } \Lambda_{KL} \ge 0, \\ \Lambda_{KL}(u_K - u_L) \min\limits_{u\in[u_K,u_L]} a(u), & \text{if } \Lambda_{KL} < 0, \\ \Lambda_{KL}(u_K - u_L) \max\limits_{u\in[u_L,u_K]} a(u), & \text{if } \Lambda_{KL} \ge 0, \\ \Lambda_{KL}(u_K - u_L) \min\limits_{u\in[u_L,u_K]} a(u), & \text{if } \Lambda_{KL} < 0. \end{cases}$$

Returning to the first equation of system (4.1). The discretization of the initial data u_K^0, $K \in \mathcal{V}$ is defined by

$$u_{\mathcal{M}}^0(\mathbf{x}) = u_K^0 = \frac{1}{m_K}\int_{\omega_K} u_0(\mathbf{y})\, \mathrm{d}\mathbf{y}, \qquad \forall \mathbf{x} \in \omega_K, \tag{4.10}$$

$$v_{\mathcal{M}}^0(\mathbf{x}) = v_K^0 = \frac{1}{m_K}\int_{\omega_K} v_0(\mathbf{y})\, \mathrm{d}\mathbf{y}, \qquad \forall \mathbf{x} \in \omega_K, \tag{4.11}$$

and for all $K \in \mathcal{V}$, and $n \in \{0, \ldots, N\}$, we define the discretization of the diffusive term by

$$\sum_{\sigma_{KL}\in\mathcal{E}_K} a_{KL}^{n+1} \Lambda_{KL}\left(u_K^{n+1} - u_L^{n+1}\right),$$

where,

$$a_{KL}^{n+1} = \begin{cases} \max\limits_{u\in I_{KL}^{n+1}} a(u) & \text{if } \Lambda_{KL} \ge 0, \\ \min\limits_{u\in I_{KL}^{n+1}} a(u) & \text{if } \Lambda_{KL} \le 0, \end{cases} \tag{4.12}$$

and I_{KL}^{n+1} denotes the interval defined by

$$I_{KL}^{n+1} = \begin{cases} \left[u_K^{n+1}, u_L^{n+1}\right] & \text{if } u_K^{n+1} \leq u_L^{n+1}, \\ \left[u_L^{n+1}, u_K^{n+1}\right] & \text{otherwise.} \end{cases}$$

Let us focus on the discretization of the convective term, and recall that the function $\chi(u)$ is defined to be the product of the continuous functions $\mu(u)$ and $a(u)$. To handle the discretization of the convective term in order to obtain a robust and stable scheme, we perform a nonclassical upwind finite volume scheme which consists of considering an upwind scheme for the function $\mu(u)$ according to the discrete gradient of v, and an upwind finite volume scheme for the function $a(u)$ with respect to u. These choices of discretization are crucial to ensure the discrete maximum principle as well as the energy estimates on the approximate solutions.

We are now in a position to introduce what we call *nonlinear control volume finite element (CVFE) scheme*. For all $K \in \mathcal{V}$, and all $n \in \{0, \ldots, N\}$,

$$\begin{aligned} \frac{u_K^{n+1} - u_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right) \\ - \sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL} \mu_{KL}^{n+1} a_{KL}^{n+1} \left(v_K^{n+1} - v_L^{n+1}\right) = f\left(u_K^{n+1}\right) m_K, \end{aligned} \tag{4.13}$$

where the transmissibility coefficients Λ_{KL} and D_{KL} are given by equality (4.8), and μ_{KL}^{n+1} denotes an approximation of $\mu(u)$ on the interfaces of ω_K with respect to the discrete gradient of v. The term $\Lambda_{KL}\mu_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right)$ represents a numerical flux function computed at $\left(u_K^{n+1}, u_L^{n+1}, \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right)\right)$. In the general case, a numerical flux function G of arguments $(a, b, c) \in \mathbb{R}^3$ is required to satisfy the following properties :

$$\begin{cases} \text{(a) } G(\cdot, b, c) \text{ is nondecreasing for all } b, c \in \mathbb{R}, \\ \quad \text{and } G(a, \cdot, c) \text{ is nonincreasing for all } a, c \in \mathbb{R}; \\ \text{(b) } G(a, b, c) = -G(b, a, -c) \text{ for all } a, b, c \in \mathbb{R}; \\ \text{(c) } G(a, a, c) = \mu(a)\, c \text{ for all } a, c \in \mathbb{R}. \end{cases} \tag{4.14}$$

We give here two examples on the construction of μ_{KL}^{n+1} such that the numerical flux function satisfies properties (4.14). The first example consists of taking the Engquist-Osher scheme [65] and the second example consists of taking the Godunov scheme (see e.g. [32, 40]).

- $\mu_{KL}^{n+1} = \begin{cases} \mu_\downarrow\left(u_K^{n+1}\right) + \mu_\uparrow\left(u_L^{n+1}\right), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \mu_\uparrow\left(u_K^{n+1}\right) + \mu_\downarrow\left(u_L^{n+1}\right), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0. \end{cases}$
  The functions $\mu_\uparrow$ and $\mu_\downarrow$ are given by
  $$\mu_\uparrow(z) := \int_0^z \left(\mu'(s)\right)^+ \mathrm{d}s, \qquad \mu_\downarrow(z) := -\int_0^z \left(\mu'(s)\right)^- \mathrm{d}s.$$
  Herein, $s^+ = \max(s, 0)$ and $s^- = \max(-s, 0)$.

- $\mu_{KL}^{n+1} = \begin{cases} \max\limits_{[u_K^{n+1}, u_L^{n+1}]} \mu(u), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \min\limits_{[u_L^{n+1}, u_K^{n+1}]} \mu(u), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0, \\ \max\limits_{[u_L^{n+1}, u_K^{n+1}]} \mu(u), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0, \\ \min\limits_{[u_K^{n+1}, u_L^{n+1}]} \mu(u), & \text{if } \Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) < 0. \end{cases}$

Remark 3. In view of properties (4.14) of the numerical flux function, one can deduce that the approximate numerical flux function $\mu_{KL}^{n+1} = z\left(u_K^{n+1}, u_L^{n+1}\right)$ must be nonincreasing (resp. nondecreasing) with respect to u_K^{n+1} and nondecreasing (resp. nonincreasing) with respect to u_L^{n+1} when $\Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \geq 0$ (resp. $\Lambda_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) \leq 0$). We also have, using property (4.14)(c), that $\mu_{KK}^{n+1} = z\left(u_K^{n+1}, u_K^{n+1}\right) = \mu\left(u_K^{n+1}\right)$.

4.4.2 Discretization of the second equation of system (4.1)

We now focus on the discretization of the second equation of system (4.1). We note that a classical discretization of this equation is given by the following form

$$m_K \frac{v_K^{n+1} - v_K^n}{\Delta t} + \sum_{\sigma_{KL} \in \mathcal{E}_K} D_{KL}\left(v_K^{n+1} - v_L^{n+1}\right) = m_K\left(\alpha u_K^n - \beta v_K^{n+1}\right). \tag{4.15}$$

However, we have seen in chapter 3 that this discretization does not guaranty the discrete maximum principle without any restriction on the transmissibility coefficients, for instance, one can get the discrete maximum principle by assuming that all the transmissibility coefficients D_{KL} are nonnegative (e.g. see [50]).
Here, we propose a numerical discretization in order to ensure the discrete maximum principle without any restriction on the transmissibility coefficients. To do this, we introduce the following set of functions : $\eta(v)$, $p(v)$, $\Gamma(v)$ and $\phi(v)$ defined by

$$\eta(v) = \max(0, \min(v, 1)), \tag{4.16}$$

$$p(v) = \int_1^v \frac{1}{\eta(s)}\,\mathrm{d}s = \begin{cases} \ln(v) & \text{if } v \in (0,1), \\ v - 1 & \text{if } v \geq 1, \end{cases} \tag{4.17}$$

$$\Gamma(v) = \int_1^v p(s)\,\mathrm{d}s = \begin{cases} v\ln(v) - v + 1 & \text{if } v \in [0,1), \\ \frac{(v-1)^2}{2} & \text{if } v \geq 1, \end{cases} \tag{4.18}$$

$$\phi(v) = \int_0^v \frac{1}{\sqrt{\eta(s)}}\,\mathrm{d}s = \begin{cases} \frac{\sqrt{v}-1}{2} & \text{if } v \in [0,1), \\ v - 1 & \text{if } v \geq 1. \end{cases} \tag{4.19}$$

In the sequel, we adopt the convention

$$\eta(v)p(v) = 0 \quad \text{if } v \leq 0. \tag{4.20}$$

We mention the discretization of the second equation of (4.1) ; specifically, we have

$$m_K \frac{v_K^{n+1} - v_K^n}{\Delta t} - \sum_{\sigma_{KL} \in \mathcal{E}_K} D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right) - p\left(v_L^{n+1}\right)\right) = m_K\left(\alpha u_K^n - \beta v_K^{n+1}\right), \tag{4.21}$$

where, denoting by $J_{KL}^{n+1} = [\min(v_K^{n+1}, v_L^{n+1}), \max(v_K^{n+1}, v_L^{n+1})]$, we have set

$$\eta_{KL}^{n+1} = \begin{cases} \max_{s \in J_{KL}^{n+1}} \eta(s) & \text{if } D_{KL} \geq 0, \\ \min_{s \in J_{KL}^{n+1}} \eta(s) & \text{if } D_{KL} < 0. \end{cases}$$

Taking into account properties (4.14) of the numerical flux function G, one can deduce that this scheme, whose construction is based on finite elements for the diffusive term and an upstream finite volume for the convective term, can be interpreted as a finite volume scheme. Indeed, denoting by

$$F_{KL}^{n+1} = \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right) - \Lambda_{KL}\mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right),$$
$$\Phi_{KL}^{n+1} = D_{KL}\eta_{KL}^{n+1}\left(p\left(v_K^{n+1}\right) - p\left(v_L^{n+1}\right)\right).$$

We have $a_{KL}^{n+1} = a_{LK}^{n+1}$, $\eta_{KL}^{n+1} = \eta_{LK}^{n+1}$, and $\mu_{KL}^{n+1} = \mu_{LK}^{n+1}$. Thus, we obtain the locally conservative form

$$\begin{cases} F_{KL}^{n+1} + F_{LK}^{n+1} = 0 = \Phi_{KL}^{n+1} + \Phi_{LK}^{n+1}, & \text{for all } \sigma_{KL} \in \mathcal{E}, \\ m_K \dfrac{u_K^{n+1} - u_K^n}{\Delta t} + \displaystyle\sum_{\sigma_{KL} \in \mathcal{E}_K} F_{KL}^{n+1} = f\left(u_K^{n+1}\right) m_K, & \text{for all } K \in \mathcal{V}, \\ m_K \dfrac{v_K^{n+1} - v_K^n}{\Delta t} + \displaystyle\sum_{\sigma_{KL} \in \mathcal{E}_K} \Phi_{KL}^{n+1} = g\left(u_K^n, v_K^{n+1}\right) m_K, & \text{for all } K \in \mathcal{V}. \end{cases}$$

4.4.3 Main result

Let $(\mathcal{T}_m)_{m\geq 1}$ be a sequence of triangulations of Ω such that

$$h_m = \max_{T \in \mathcal{T}_m} \operatorname{diam}(T) \to 0 \text{ as } m \to \infty.$$

We assume that there exists a constant $\theta > 0$ such that

$$\theta_{\mathcal{T}_m} \leq \theta, \quad \forall m \geq 1.$$

As before, a sequence of dual meshes $(\mathcal{M}_m)_{m\geq 1}$ is given.

Let $(N_m)_m$ be an increasing sequence of integers, then we define the corresponding sequence of time steps $(\Delta t_m)_m$ such that $\Delta t_m \to 0$ as $m \to \infty$. The intention of this chapter is to prove the following main result.

Theorem 4.2. *Let $(u_{\mathcal{M}_m,\Delta t_m}, v_{\mathcal{M}_m,\Delta t_m})_m$ be a sequence of solutions to the scheme (4.13)–(4.21), such that $0 \leq u_{\mathcal{M}_m,\Delta t_m} \leq 1$ and $0 \leq v_{\mathcal{M}_m,\Delta t_m}$ for almost everywhere in Q_{t_f}, then*

$$u_{\mathcal{M}_m,\Delta t_m} \to u \text{ and } v_{\mathcal{M}_m,\Delta t_m} \to v \quad \text{a.e. in } Q_{t_\mathrm{f}} \text{ as } m \to \infty,$$

where the couple (u, v) is a weak solution to the system (4.1)–(4.3) in the sense of Definition 4.1.

4.5 Discrete properties, a priori estimates and existence

In this section, we first bring up some preliminary lemmas presented by Cances and Guichard [16], that we reproduce them here for clarity. We then establish the discrete maximum principle which is the basis to the analysis that we are going to perform. Next, we carry out the *a priori* estimates necessary to prove the existence of a discrete solution to the discrete problem (4.13)–(4.21). The maximum principle and the *a priori* estimates are crucial to prove the convergence of the scheme (4.13)–(4.21) towards the weak solution.

Lemma 4.3. *Let $\left(u_K^{n+1}\right)_{K,n} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$ (resp. $\left(v_K^{n+1}\right)_{K,n} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$), then denoting by $\xi_{\mathcal{T},\Delta t}$ (resp. $\phi_{\mathcal{T},\Delta t}$) the unique function of $\mathcal{H}_{\mathcal{T},\Delta t}$ with nodal values $\left(\xi\left(u_K^{n+1}\right)\right)_{K,n} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$ (resp. $\left(\phi\left(v_K^{n+1}\right)\right) \in \mathbb{R}^{(N+1)\#\mathcal{V}}$), one has*

$$\begin{aligned} &\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2 \\ &\qquad \geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} \Lambda_{KL} \left(\xi\left(u_K^{n+1}\right) - \xi\left(u_L^{n+1}\right)\right)^2 = \iint_{Q_{t_\mathrm{f}}} \Lambda \nabla \xi_{\mathcal{T},\Delta t} \cdot \nabla \xi_{\mathcal{T},\Delta t} \mathrm{d}\mathbf{x}\mathrm{d}t. \end{aligned} \tag{4.22}$$

and

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1})-p(v_L^{n+1})\right)^2$$
$$\geq\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}\left(\phi\left(v_K^{n+1}\right)-\phi\left(v_L^{n+1}\right)\right)^2=\iint_{Q_{t_f}}D\nabla\phi_{\mathcal{T},\Delta t}\cdot\nabla\phi_{\mathcal{T},\Delta t}\mathrm{d}\mathbf{x}\mathrm{d}t. \tag{4.23}$$

Proof. Using the mean value theorem since ξ is a differentiable function, one can deduce the existence of $c\in\overset{\circ}{I}{}_{KL}^{n+1}$, such that

$$\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2=a\left(c\right)\left(u_K^{n+1}-u_L^{n+1}\right)^2\qquad\forall\sigma_{KL}\in\mathcal{E}.$$

Multiplying this equality by Λ_{KL}, one gets

$$\Lambda_{KL}\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2\leq\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2,$$

for the case where $\Lambda_{KL}\geq 0$ since a_{KL}^{n+1} is the maximum of $a\left(u\right)$ on the interval I_{KL}^{n+1}. On the other hand, when $\Lambda_{KL}\leq 0$, one has that a_{KL}^{n+1} is the minimum of $a\left(u\right)$ on the interval I_{KL}^{n+1}, thus $\Lambda_{KL}a\left(c\right)\leq\Lambda_{KL}a_{KL}^{n+1}$ and property (4.22) holds. The proof of inequality (4.23) is similar. This ends the proof of the lemma. □

Let $T\in\mathcal{T}$, and let $(K,\,L)\in\mathcal{V}^2$, we denote by

$$\lambda_{KL}^T:=-\int_T\Lambda\left(\mathbf{x}\right)\nabla\varphi_K\left(\mathbf{x}\right)\cdot\nabla\varphi_L\left(\mathbf{x}\right)\mathrm{d}\mathbf{x}=\lambda_{L,K}^T.$$

In particular, one has $\Lambda_{KL}=-\sum_{T\in\mathcal{T}}\int_T\Lambda\nabla\varphi_K\cdot\nabla\varphi_L\mathrm{d}\mathbf{x}=\sum_{T\in\mathcal{T}}\lambda_{KL}^T,$ for all $\sigma_{KL}\in\mathcal{E}$.

Lemma 4.4. *Let $\Psi_{\mathcal{T}}=\sum_{K\in\mathcal{V}}\psi_K\varphi_K\in\mathcal{H}_{\mathcal{T}}$, then there exists a quantity C_0 depending only on Λ, $\theta_{\mathcal{T}}$ such that*

$$\sum_{\sigma_{KL}\in\mathcal{E}}\sum_{T\in\mathcal{T}}\left|\lambda_{KL}^T\right|\left(\psi_K-\psi_L\right)^2\leq C_0\int_\Omega\Lambda\nabla\Psi_{\mathcal{T}}\cdot\nabla\Psi_{\mathcal{T}}\mathrm{d}\mathbf{x}. \tag{4.24}$$

Proof. In the proof below, unless specified, C denotes a generic quantity depending only on Λ and $\theta_{\mathcal{T}}$.
In order to prove inequality (4.24), it suffices to prove that

$$\sum_{\sigma_{KL}\in\mathcal{E}}\sum_{T\in\mathcal{T}}\left|\lambda_{KL}^T\right|\left(\psi_K-\psi_L\right)^2\leq C\left\|\nabla\Psi_{\mathcal{T}}\right\|_{(L^2(\Omega))^2}^2,$$

since we have, using the assumption on Λ, that

$$\left\|\nabla\Psi_{\mathcal{T}}\right\|_{L^2(\Omega)}^2=\int_\Omega\nabla\Psi_{\mathcal{T}}\cdot\nabla\Psi_{\mathcal{T}}\mathrm{d}\mathbf{x}\leq\frac{1}{\Lambda_-}\int_\Omega\Lambda\nabla\Psi_{\mathcal{T}}\cdot\nabla\Psi_{\mathcal{T}}\mathrm{d}\mathbf{x}.$$

Thanks to the Cauchy-Schwarz inequality, one has

$$\left|\lambda_{KL}^T\right|\leq\Lambda_-\left\|\nabla\varphi_K\right\|_{(L^2(T))^2}\left\|\nabla\varphi_L\right\|_{(L^2(T))^2}. \tag{4.25}$$

Now, using a classical inequality stemming from finite element properties (see e.g. [27, 11]), one gets

$$\|\nabla\varphi_K\|^2_{L^2(T)} \leq c\theta_{\mathcal{T}} \frac{|T|}{(h_T)^2}, \quad \forall K \in \mathcal{V},\ \forall T \in \mathcal{T}, \tag{4.26}$$

where c is an absolute constant.
Plugging inequality (4.26) into (4.25), one can conclude that

$$\left|\lambda_{KL}^T\right| \leq C \frac{|T|}{(h_T)^2}, \quad \forall T \in \mathcal{T},\ \forall \sigma_{KL} \in \mathcal{E}_{\mathcal{T}}. \tag{4.27}$$

This implies that

$$\sum_{\sigma_{KL}\in\mathcal{E}_T} \left|\lambda_{KL}^T\right| (\psi_K - \psi_L)^2 \leq C \frac{|T|}{(h_T)^2} \sum_{\sigma_{KL}\in\mathcal{E}_T} (\psi_K - \psi_L)^2.$$

Finally, it follows form the analysis carried out for example in [11] that for all $T \in \mathcal{T}$ with K, L, and M being its vertices

$$\frac{|T|}{(h_T)^2}\left((\psi_K - \psi_L)^2 + (\psi_K - \psi_M)^2 + (\psi_L - \psi_M)^2\right) \leq C \left\|\nabla\Psi_{\mathcal{T}}\right\|^2_{(L^2(T))^2}.$$

Since $\sigma_{KL} \in \mathcal{E}$ is shared by at most two triangles, one has

$$\sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}\in\mathcal{E}_T} \left|\lambda_{KL}^T\right| (\psi_K - \psi_L)^2 \leq C \sum_{T\in\mathcal{T}} \left\|\nabla\Psi_{\mathcal{T}}\right\|^2_{L^2(T)} = C \left\|\nabla\Psi_{\mathcal{T}}\right\|^2_{(L^2(\Omega))^2},$$

this concludes the proof of the lemma. □

Lemma 4.5. *There exists a quantity C_1 depending only on Λ and $\theta_{\mathcal{T}}$ such that*

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |\Lambda_{KL}|\, a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2 \leq C_1 \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2. \tag{4.28}$$

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |D_{KL}|\, \eta_{KL}^{n+1} \left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2 \leq C_1 \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} D_{KL} \eta_{KL}^{n+1} \left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2. \tag{4.29}$$

Proof. We denote by $\mathcal{E}^- := \{\sigma_{KL} \in \mathcal{E};\ \Lambda_{KL} < 0\}$, then since $|\mathbf{x}| = \mathbf{x} + 2\mathbf{x}^-$, $\mathbf{x}^- = \max(-\mathbf{x}, 0)$, one has

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |\Lambda_{KL}|\, a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2 = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2 + 2\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}^-} |\Lambda_{KL}|\, a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2.$$

Now, from the definition (4.8) of a_{KL}^{n+1}, there exists $c \in \overset{\circ}{I}{}_{KL}^{n+1}$ such that

$$\left(\xi\left(u_K^{n+1}\right) - \xi\left(u_L^{n+1}\right)\right)^2 = a(c)\left(u_K^{n+1} - u_L^{n+1}\right)^2 \geq a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right)^2, \quad \forall \sigma_{KL} \in \mathcal{E}^-.$$

Therefore, it yields

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|\Lambda_{KL}|\,a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2\leq\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2 + 2\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|\Lambda_{KL}|\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2. \tag{4.30}$$

Lemma 4.4 ensures the existence of a quantity $C_0>0\,(=C_0\,(\Lambda,\theta_{\mathcal{T}}))$ such that

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|\Lambda_{KL}|\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2 \leq\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\sum_{T\in\mathcal{T}}|\lambda_{KL}|\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2\leq C_0\int_{Q_t}\Lambda\nabla\xi_{\mathcal{T},\Delta t}\cdot\nabla\xi_{\mathcal{T},\Delta t}\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

and from lemma 4.3, we deduce that

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|\Lambda_{KL}|\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)^2\leq C_0\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2. \tag{4.31}$$

Plugging estimate (4.31) into estimate (4.30) , then estimate (4.28) holds with $C_1=1+2C_0$. The proof of (4.29) is similar. This ends the proof of the lemma. □

4.5.1 Discrete maximum principle

Lemma 4.6. *Let $\left(u_K^{n+1}\right)_{K\in\mathcal{V},\,n\in\{0,\ldots,N\}}$ be a solution to the CVFE scheme (4.13). Then, for all $K\in\mathcal{V}_h$, and all $n\in\{0,\ldots,N+1\}$, we have $0\leq u_K^n\leq 1$ and $v_K^n\geq 0$.*

Proof. We want to show this property using the induction on n. The property is true for $n=0$ thanks to the definition (4.10) of u_K^0 and to the assumption on u_0. Now, assume that the claim is true up to time step n. Consider a dual control volume ω_K such that $u_K^{n+1}=\min\limits_{L\in\mathcal{V}}\{u_L^{n+1}\}$, we want to show that $u_K^{n+1}\geq 0$ i.e. $\left(u_K^{n+1}\right)^-=0$.
Multiplying the scheme (4.13) by $-\left(u_K^{n+1}\right)^-$, one has

$$-m_K\frac{u_K^{n+1}-u_K^n}{\Delta t}\left(u_K^{n+1}\right)^- -\sum_{\sigma_{KL}\in\mathcal{E}_K}\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)\left(u_K^{n+1}\right)^- + \sum_{\sigma_{KL}\in\mathcal{E}_K}\Lambda_{KL}\mu_{KL}^{n+1}a_{KL}^{n+1}\left(v_K^{n+1}-v_L^{n+1}\right)\left(u_K^{n+1}\right)^- =-m_K f\left(u_K^{n+1}\right)\left(u_K^{n+1}\right)^-, \tag{4.32}$$

for which, we use the extension by $f\,(0)\geq 0$ of the function f for $u\leq 0$, then the right hand side of equation (4.32) is less or equal to zero.
In view of definition (4.12) of a_{KL}^{n+1}, and of the fact that $a\,(u)=0$ for every $u\leq 0$, one has

$$a_{KL}^{n+1}\left(u_K^{n+1}\right)^-=0,\qquad\text{if }\Lambda_{KL}\leq 0.$$

Indeed, assume that $u_K^{n+1} \geq 0$ then $\left(u_K^{n+1}\right)^- = 0$ and $a_{KL}^{n+1}\left(u_K^{n+1}\right)^- = 0$. Nevertheless, assume that $u_K^{n+1} < 0$ then $a_{KL}^{n+1} = \min\limits_{u \in I_{KL}^{n+1}} a(u) = 0$ if $\Lambda_{KL} \leq 0$.
As a consequence, the second term in the left hand side of equation (4.32) reads to

$$-\sum_{\sigma_{KL} \in \mathcal{E}_K} a_{KL}^{n+1} \left(\Lambda_{KL}\right)^+ \left(u_K^{n+1} - u_L^{n+1}\right)\left(u_K^{n+1}\right)^- \geq 0,$$

Let us now focus on the third term of equation (4.32), and denote by $\mathcal{A}$ this term. Using the fact that $a_{KL}^{n+1}\left(u_K^{n+1}\right)^- = 0$ for $\Lambda_{KL} \leq 0$, then $\mathcal{A}$ rewrites

$$\begin{aligned}\mathcal{A} = &\sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL}^+ \mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right)^+ \left(u_K^{n+1}\right)^- \\ &- \sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL}^+ \mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right)^- \left(u_K^{n+1}\right)^-.\end{aligned}$$

The second term of $\mathcal{A}$ is nonpositive, but in view of Remark 3 on the approximation μ_{KL}^{n+1} and in view of the extension by zero of the function μ for $u \leq 0$ (since $\mu(0) = 0$), one can deduce that

$$\mu_{KL}^{n+1}\Lambda_{KL}^+\left(v_K^{n+1} - v_L^{n+1}\right)^-\left(u_K^{n+1}\right)^- \leq \mu\left(u_K^{n+1}\right)\Lambda_{KL}^+\left(v_K^{n+1} - v_L^{n+1}\right)^-\left(u_K^{n+1}\right)^- = 0,$$

thus, the second term of $\mathcal{A}$ is equal to zero and consequently $\mathcal{A} \geq 0$ since the first term of $\mathcal{A}$ is nonnegative.

Finally, we use the identity $u_K^{n+1} = \left(u_K^{n+1}\right)^+ - \left(u_K^{n+1}\right)^-$ and the nonnegativity of u_K^n, one can deduce from equation (4.32) that $\left(u_K^{n+1}\right)^- = 0$. According to the choice of the dual control volume ω_K, then $\min\limits_{L \in \mathcal{V}}\{u_L^{n+1}\}$ is non-negative. Consequently,

$$u_K^n \geq 0, \quad \forall K \in \mathcal{V}, \text{ and all } n \in \{0, \ldots, N+1\}.$$

In order to prove by induction that $u_K^n \leq 1$, $\forall K \in \mathcal{V}$, $\forall n \in \{0, \ldots, N+1\}$, we proceed in the same way as before, so that we assume that the claim $u_K^n \leq 1$ is true for all $K \in \mathcal{V}$ and we consider a dual control volume ω_K such that $u_K^{n+1} = \max\limits_{L \in \mathcal{V}}\{u_L^{n+1}\}$, we want to show that $u_K^{n+1} \leq 1$.

For the mentioned claim, we multiply equation (4.13) by $\left(u_K^{n+1} - 1\right)^+$, one gets

$$\begin{aligned}m_K \frac{u_K^{n+1} - u_K^n}{\Delta t}\left(u_K^{n+1} - 1\right)^+ + &\sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right)\left(u_K^{n+1} - 1\right)^+ \\ &- \sum_{\sigma_{KL} \in \mathcal{E}_K} \Lambda_{KL}\mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1} - v_L^{n+1}\right)\left(u_K^{n+1} - 1\right)^+ = 0,\end{aligned} \tag{4.33}$$

for which, we have use the extension by $f(1) \leq 0$ of the function f for $u \geq 1$.
In view of definition (4.12) of a_{KL}^{n+1}, and of the fact that $a(u) = 0$ for every $u \geq 1$, one has

$$a_{KL}^{n+1}\left(u_K^{n+1} - 1\right)^+ = 0, \qquad \text{if } \Lambda_{KL} \leq 0.$$

the second term in the left hand side of equation (4.33) reads to

$$\sum_{\sigma_{KL} \in \mathcal{E}_K} a_{KL}^{n+1}\left(\Lambda_{KL}\right)^+\left(u_K^{n+1} - u_L^{n+1}\right)\left(u_K^{n+1} - 1\right)^+ \geq 0. \tag{4.34}$$

Using estimate (4.34) and denoting by $\mathcal{A}_1$ the third term of equation (4.33), one has

$$\mathcal{A}_1 \geq -\sum_{\sigma_{KL}\in\mathcal{E}_K} (\Lambda_{KL})^+ \mu_{KL}^{n+1} a_{KL}^{n+1} \left(v_K^{n+1} - v_L^{n+1}\right)^+ \left(u_K^{n+1} - 1\right)^+. \tag{4.35}$$

The right hand side of inequality (4.35) is nonpositive, but in view of Remark 3 on the approximation μ_{KL}^{n+1} and in view of the extension by zero of the function μ for $u \geq 1$ (since $\mu(1) = 0$), one can deduce that

$$\mu_{KL}^{n+1}\Lambda_{KL}^+ \left(v_K^{n+1} - v_L^{n+1}\right)^+ \left(u_K^{n+1} - 1\right)^+ \leq \mu\left(u_K^{n+1}\right)\Lambda_{KL}^+ \left(v_K^{n+1} - v_L^{n+1}\right)^+ \left(u_K^{n+1} - 1\right)^+ = 0,$$

and consequently, one gets that $\mathcal{A}_1$ is nonnegative.
Using the identity $\left(u_K^{n+1} - 1\right) = \left(u_K^{n+1} - 1\right)^+ - \left(u_K^{n+1} - 1\right)^-$ and that $u_K^n \leq 1$, one can deduce from equation (4.33) that $\left(u_K^{n+1} - 1\right)^+ = 0$. According to the choice of the dual control volume ω_K, then $\max_{L\in\mathcal{V}}\{u_L^{n+1}\}$ is non-negative. Consequently,

$$u_K^n \leq 1, \quad \forall K \in \mathcal{V}, \text{ and all } n \in \{0, \ldots, N+1\}.$$

Let us prove now that $v_K^{n+1} \geq 0$ for all $K \in \mathcal{V}$. Let $K_m \in \mathcal{V}$ be such that $v_{K_m}^{n+1} = \min_{K\in\mathcal{V}} v_K^{n+1}$, and assume that $v_{K_m}^{n+1} \leq 0$. Thanks to the convention (4.20), we can claim that

$$D_{K_mL}\eta_{K_mL}^{n+1}(p(v_{K_m}^{n+1}) - p(v_L^{n+1})) = 0 \quad \text{if } D_{K_mL} \leq 0,$$

and that

$$D_{K_mL}\eta_{K_mL}^{n+1}(p(v_{K_m}^{n+1}) - p(v_L^{n+1})) \leq 0 \quad \text{if } D_{K_mL} \geq 0.$$

Therefore, the scheme (4.21) together with the positivity of $u_{K_m}^n$ yields

$$v_{K_m}^{n+1} \geq \frac{v_{K_m}^n}{1+\beta\Delta t} \geq 0.$$

This achieves the proof of Lemma 4.6. □

4.5.2 Entropy estimates on $v_{\mathcal{M},\Delta t}$

Lemma 4.7. *There exists $C > 0$ depending only on $\|v_0\|_{L^2(\Omega)}$, Ω, t_f, α and β such that, for all $n^\star \in \{0, \ldots, N\}$, one has*

$$\sum_{K\in\mathcal{V}} m_K\Gamma(v_K^{n^\star+1}) + \sum_{n=0}^{n^\star}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2 \leq C.$$

Proof. It follows from Jensen's inequality – recall that Γ is convex – that

$$\sum_{K\in\mathcal{V}} m_K\Gamma(v_K^0) \leq \int_\Omega \Gamma(v_0(\mathbf{x}))\mathrm{d}\mathbf{x}.$$

Since $\Gamma(v) \leq (v-1)^2$ for all $v \geq 0$, we obtain that

$$\sum_{K\in\mathcal{V}} m_K\Gamma(v_K^0) \leq \int_\Omega (v_0(\mathbf{x}) - 1)^2\mathrm{d}\mathbf{x} \leq C. \tag{4.36}$$

Multiplying the scheme (4.21) by $p(v_K^{n+1})\Delta t$ and summing of $K \in \mathcal{V}$ and $n = 0, \ldots, n^\star$ provides

$$\mathcal{A} + \mathcal{B} = \mathcal{C}, \tag{4.37}$$

where we have set

$$\mathcal{A} = \sum_{n=0}^{n^\star} \sum_{K\in\mathcal{V}} m_K(v_K^{n+1} - v_K^n)p(v_K^{n+1}),$$
$$\mathcal{B} = \sum_{n=0}^{n^\star} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2,$$
$$\mathcal{C} = \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} m_K(\alpha u_K^n - \beta v_K^{n+1})p(v_K^{n+1}).$$

Since, thanks to Lemma 4.6, u_K^n is nonnegative for all $K \in \mathcal{V}$ and all $n \geq 0$, and since $p(v) \leq (v-1)$ for all $v \geq 0$ (with the convention $p(0) = -\infty$), one has

$$\alpha u_K^n p(v_K^{n+1}) \leq \alpha u_K^n (v_K^{n+1} - 1).$$

On the other hand, there exists an absolute constant $c^\star$ such that $vp(v) \geq (v-1)^2 - c^\star$ for all $v \geq 0$. Therefore,

$$\beta v_K^{n+1} p(v_K^{n+1}) \geq \beta(v_K^{n+1} - 1)^2 - c^\star.$$

As a consequence, we obtain that

$$\mathcal{C} \leq t_{\rm f}|\Omega|c^\star + \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} m_K\left(\alpha u_K^n(v_K^{n+1} - 1) - \beta(v_K^{n+1} - 1)^2\right).$$

Using the weighted Young's inequality $\alpha ab \leq \beta b^2 + \frac{\alpha^2}{4\beta}a^2$ for all $(a,b) \in \mathbb{R}^2$ provides

$$\alpha u_K^n(v_K^{n+1} - 1) - \beta(v_K^{n+1} - 1)^2 \leq \frac{\alpha^2}{4\beta}u_K^n \leq \frac{\alpha^2}{4\beta}$$

thanks to Lemma 4.6. Hence, we obtain that

$$\mathcal{C} \leq t_{\rm f}|\Omega|\left(c^\star + \frac{\alpha^2}{4\beta}\right). \tag{4.38}$$

The function p being increasing, an elementary convexity inequality provides that

$$(a-b)p(a) \geq \Gamma(a) - \Gamma(b), \quad \forall (a,b) \in (\mathbb{R}_+)^2,$$

ensuring that

$$\mathcal{A} \geq \sum_{n=0}^{n^\star} \sum_{K\in\mathcal{V}} m_K\left(\Gamma(v_K^{n+1}) - \Gamma(v_K^n)\right) = \sum_{K\in\mathcal{V}} m_K\left(\Gamma(v_K^{n^\star+1}) - \Gamma(v_K^0)\right). \tag{4.39}$$

Using (4.38), (4.39) and (4.36) in (4.37) concludes the proof of Lemma 4.7. □

Lemma 4.8. *There exists C depending only on $\|v_0\|_{L^2(\Omega)}$, Ω, $t_{\rm f}$, α, β, $D_\pm$, Λ_+ and $\theta_{\mathcal{T}}$, such that*

$$\iint_{Q_{t_{\rm f}}} \Lambda(\mathbf{x})\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\cdot\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\mathrm{d}\mathbf{x}\,\mathrm{d}t = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL}(v_K^{n+1} - v_L^{n+1})^2 \leq C. \tag{4.40}$$

Proof. Thanks to Lemmas 4.3 and 4.7, we know that

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}(\phi(v_K^{n+1})-\phi(v_L^{n+1}))^2\le\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}\eta_{KL}^{n+1}\left(p(v_K^{n+1})-p(v_L^{n+1})\right)^2\le C.$$

Therefore, it follows from Lemma 4.4 that

$$\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|D_{KL}|(\phi(v_K^{n+1})-\phi(v_L^{n+1}))^2\le C.$$

since $(\phi(v_K^{n+1})-\phi(v_L^{n+1}))^2\ge(v_K^{n+1}-v_L^{n+1})^2$, we obtain that

$$\begin{aligned}\iint_{Q_{t_{\rm f}}}D(\mathbf{x})\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\cdot\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\mathrm{d}\mathbf{x}\,\mathrm{d}t&=\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}D_{KL}(v_K^{n+1}-v_L^{n+1})^2\\&\le\sum_{n=0}^{N}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}|D_{KL}|(v_K^{n+1}-v_L^{n+1})^2\le C.\end{aligned}$$

It only remains to check that

$$\iint_{Q_{t_{\rm f}}}\Lambda(\mathbf{x})\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\cdot\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\mathrm{d}\mathbf{x}\,\mathrm{d}t\le\frac{\Lambda_+}{D_-}\iint_{Q_{t_{\rm f}}}D(\mathbf{x})\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\cdot\nabla v_{\mathcal{T},\Delta t}(\mathbf{x},t)\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

in order to conclude the proof of Lemma 4.8. □

Remark 4. A careful analysis allows to prove, after a more involving proof very similar to the one of Lemma 4.7, that the constant C depends neither on D_+ nor on $\theta_{\mathcal{T}}$. The cornerstone of the corresponding proof is the inequality

$$D_{KL}\eta_{KL}^{n+1}(p(v_K^{n+1})-p(v_L^{n+1}))(v_K^{n+1}-v_L^{n+1})\ge D_{KL}(v_K^{n+1}-v_L^{n+1})^2$$

that relies on the definition of η_{KL}^{n+1} and on the link between the functions η and p.

4.5.3 Energy estimates on $u_{\mathcal{M},\Delta t}$

In the following, C denotes a "generic" constant, which need not have the same value throughout the proofs.

Proposition 4.9. *Let $\left(u_K^{n+1},v_K^{n+1}\right)_{K\in\mathcal{V},\,n\in\{0,\dots,N\}}$ be a solution to the scheme (4.13)-(4.21). There exists a constant $C>0$ depending only on $\|v_0\|_{L^2(\Omega)}$, Ω, $t_{\rm f}$, α, β, Λ, D, and $\theta_{\mathcal{T}}$ such that*

$$\sum_{K\in\mathcal{V}}m_K\left(u_K^{n^\star+1}\right)^2+\sum_{n=0}^{n^\star}\Delta t\sum_{\sigma_{KL}\in\mathcal{E}}\Lambda_{KL}a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2\le C.\tag{4.41}$$

Proof. We multiply equation (4.13) by $\Delta t\,u_K^{n+1}$ and sum over $K\in\mathcal{V}$ and $n\in\{0,\dots,n^\star\}$. This yields

$$E_1+E_2+E_3=E_4,\tag{4.42}$$

where

$$
\begin{aligned}
E_1 &= \sum_{n=0}^{n^\star} \sum_{K\in\mathcal{V}} m_K \left(u_K^{n+1} - u_K^n\right) u_K^{n+1},\\
E_2 &= \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right) u_K^{n+1},\\
E_3 &= -\sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} \mu_{KL}^{n+1} a_{KL}^{n+1} \left(v_K^{n+1} - v_L^{n+1}\right) u_K^{n+1},\\
E_4 &= \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} m_K f\left(u_K^{n+1}\right) u_K^{n+1}.
\end{aligned}
$$

For the time evolution term, we use the following inequality : $(a-b)\,a \geq \dfrac{1}{2}\left(a^2-b^2\right),\ \forall a,b\in\mathbb{R}$, to get

$$
E_1 \geq \frac{1}{2}\sum_{n=0}^{n^\star}\sum_{K\in\mathcal{V}_h} m_K\left(\left(u_K^{n+1}\right)^2 - \left(u_K^n\right)^2\right) = \frac{1}{2}\sum_{K\in\mathcal{V}_h} m_K\left(\left(u_K^{n^\star+1}\right)^2 - \left(u_K^0\right)^2\right). \tag{4.43}
$$

Next, for the diffusion term, we reorganize the sum over the edges, we find

$$
\begin{aligned}
E_2 &= \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right) u_K^{n+1}\\
&= \sum_{n=0}^{n^\star} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2.
\end{aligned} \tag{4.44}
$$

Similarly, we reorganize the sum over the edges for the convection term, we obtain

$$
\begin{aligned}
E_3 &= -\sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} \mu_{KL}^{n+1} a_{KL}^{n+1} \left(v_K^{n+1} - v_L^{n+1}\right) u_K^{n+1}\\
&= -\sum_{n=0}^{n^\star} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \Lambda_{KL} \mu_{KL}^{n+1} a_{KL}^{n+1} \left(v_K^{n+1} - v_L^{n+1}\right)\left(u_K^{n+1} - u_L^{n+1}\right).
\end{aligned}
$$

Using the weighted Young inequality, we deduce

$$
\begin{aligned}
|E_3| &\leq C\sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} |\Lambda_{KL}|\, a_{KL}^{n+1} \left|v_K^{n+1} - v_L^{n+1}\right| \left|u_K^{n+1} - u_L^{n+1}\right|\\
&\leq C\sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} |\Lambda_{KL}| \left(v_K^{n+1} - v_L^{n+1}\right)^2\\
&\quad + \frac{1}{2C_1}\sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} |\Lambda_{KL}|\, a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2.
\end{aligned}
$$

Thanks to estimates (4.24) and (4.28), one has

$$|E_3| \le C \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} \left(v_K^{n+1} - v_L^{n+1}\right)^2 + \frac{1}{2} \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2.$$

Therefore, Lemma 4.8 provides that

$$|E_3| \le C + \frac{1}{2} \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}} \sum_{\sigma_{KL}\in\mathcal{E}_K} \Lambda_{KL} a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right)^2. \tag{4.45}$$

Finally, for the reaction term, since $0 \le u_K^{n+1} \le 1$ thanks to Lemma 4.6, one has

$$E_4 = \sum_{n=0}^{n^\star} \Delta t \sum_{K\in\mathcal{V}_h} m_K f\left(u_K^{n+1}\right) u_K^{n+1} \le |\Omega| \, \|f\|_{L^\infty(0,1)} \, t_{\rm f}. \tag{4.46}$$

Plugging estimates (4.43)–(4.46) into equation (4.42), one can deduce that the estimate (4.41) holds. □

4.5.4 Enhanced estimate on $v_{\mathcal{M},\Delta t}$

The goal of this section is to prove a refined estimate on $v_{\mathcal{M},\Delta t}$ inspired from [16, Lemma 3.10], claiming that either $v_{\mathcal{M},\Delta t}$ is constant equal to 0, or $v_{\mathcal{M},\Delta t} \ge r_h > 0$ for some r_h depending on the discretization parameters. The first step consists in bounding from below the $L^\infty((0,t_{\rm f});L^1(\Omega))$ norm of $v_{\mathcal{M},\Delta t}$.

Lemma 4.10. *Assume that $\int_\Omega u_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$ or $\int_\Omega v_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$, then there exists $\kappa > 0$ depending on the discretization and on the data such that*

$$\int_\Omega v_{\mathcal{M},\Delta t}(\mathbf{x},t)\mathrm{d}\mathbf{x} \ge \kappa, \quad \forall t \in [0,t_{\rm f}].$$

Proof. Summing equation (4.21) over $K \in V$ ensures that

$$\sum_{K\in\mathcal{V}} m_K(1+\beta\Delta t) v_K^{n+1} = \sum_{K\in\mathcal{V}} m_K v_K^n + \alpha\Delta t \sum_{K\in\mathcal{V}} m_K u_K^n, \quad \forall n \in \{0,\dots,N\}. \tag{4.47}$$

Assume that $v_{K_\star^n}^n > 0$ or $u_{K_\star^n}^n > 0$ for some $K_\star^n \in \mathcal{V}$, as this is the case for $n = 0$ because of the assumption on the initial data u_0 and v_0, then we deduce from (4.47) and from the non-negativity of v_K^n and u_K^n proved in Lemma 4.6 that

$$\sum_{K\in\mathcal{V}} m_K(1+\beta\Delta t) v_K^{n+1} > 0.$$

In particular, there exists $K_\star^{n+1} \in \mathcal{V}$ such that $v_{K_\star^{n+1}}^{n+1}$ is (strictly) positive and

$$\sum_{K\in\mathcal{V}} m_K v_K^{n+1} := \kappa_{n+1} > 0.$$

One concludes the proof by setting $\kappa = \min_{n=1,\dots,N+1} \kappa_n$. □

We give now the definition of D-transmissive path, which was introduced in [16, Definition 3.4].

Definition 4.11. A *D-transmissive path* w joining $K_\mathrm{i} \in \mathcal{V}$ to $K_\mathrm{f} \in \mathcal{V}$ consists in a list of vertices $(K_q)_{0\le q\le M}$ such that $K_\mathrm{i} = K_0$, $K_\mathrm{f} = K_M$, with $K_q \neq K_\ell$ if $q \neq \ell$, and such that $\sigma_{K_qK_{q+1}} \in \mathcal{E}$ with $D_{K_qK_{q+1}} > 0$ for all $q \in \{0,\dots,M-1\}$. We denote by $\mathcal{W}(K_\mathrm{i}, K_\mathrm{f})$ the set of the transmissive path joining $K_\mathrm{i} \in \mathcal{V}$ to $K_\mathrm{f} \in \mathcal{V}$.

We now state a result which is proved in [16, Lemma 3.5].

Lemma 4.12. *For all $(K_\mathrm{i}, K_\mathrm{f}) \in \mathcal{V}^2$ there exists a transmissive path $\mathrm{w} \in \mathcal{W}(K_\mathrm{i}, K_\mathrm{f})$.*

We have now introduced all the necessary tools for proving the main result of this section.

Lemma 4.13. *Assume that $\int_\Omega u_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$ or $\int_\Omega v_0(\mathbf{x})\mathrm{d}\mathbf{x} > 0$, then there exists $r_h > 0$ depending on the data as well as on the mesh $\mathcal{T}$ and Δt such that*

$$v_K^{n+1} \ge r_h, \quad \forall K \in \mathcal{V}, \forall n \in \{0,\dots,N\}. \tag{4.48}$$

Proof. Thanks to Lemma 4.10, we know that there exists K_i such that $v_{K_\mathrm{i}}^{n+1} > 0$. Let $K_\mathrm{f} \in \mathcal{V}$, then there exists a D-transmissive path $\mathrm{w} = (K_q)_{0\le q\le M} \in \mathcal{W}(K_\mathrm{i}, K_\mathrm{f})$ thanks to Lemma 4.12, with $K_0 = K_\mathrm{i}$ and $K_M = K_\mathrm{f}$.

Thanks to Lemmas 4.5 and 4.7, we know that there exists C such that

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |D_{KL}|\, \eta_{KL}^{n+1} \left(p(v_K^{n+1}) - p(v_L^{n+1})\right)^2 \le C.$$

In particular, this ensures that

$$D_{K_qK_{q+1}} \eta_{K_qK_{q+1}}^{n+1} \left(p(v_{K_q}^{n+1}) - p(v_{q+1}^{n+1})\right)^2 \le \frac{C}{\Delta t}, \quad \forall q \in \{0,\dots,M-1\}.$$

Assume now that $v_{K_q}^{n+1} > 0$, as this is the case for $q = 0$, then $\eta_{K_qK_{q+1}}^{n+1} \ge \eta(v_{K_q}^{n+1}) > 0$. Then one has

$$\left(p(v_{K_q}^{n+1}) - p(v_{K_{q+1}}^{n+1})\right)^2 \le \frac{C}{\Delta t D_{K_qK_{q+1}} \eta_{K_qK_{q+1}}^{n+1}} < \infty. \tag{4.49}$$

Since $\lim_{v\to 0} p(v) = -\infty$, we deduce from (4.49) that $p(v_{K_{q+1}}^{n+1}) > -\infty$, hence $v_{K_{q+1}}^{n+1} > 0$. A straightforward induction provides that $v_{K_\mathrm{f}}^{n+1} > 0$, and since K_f was chosen arbitrarily, we obtain that

$$v_K^{n+1} > 0, \quad \forall K \in \mathcal{V}.$$

Since the set $\mathcal{V} \times \{0,\dots,N\}$ is finite, we can conclude that there exists r_h such that (4.48) holds. □

4.5.5 Existence of a discrete solution

Proposition 4.14. *Given $(u_K^n, v_K^n)_{K\in\mathcal{V}}$ such that $u_{\mathcal{M},\Delta t}(\cdot, n\Delta t)$ and $v_{\mathcal{M},\Delta t}(\cdot, n\Delta t)$ are nonnegative, then there exists (at least) one solution $\left(u_K^{n+1}, v_K^{n+1}\right)_{K\in\mathcal{V}}$ of the scheme (4.13),(4.21). Moreover, $u_{\mathcal{M},\Delta t}(\cdot,(n+1)\Delta t)$ and $v_{\mathcal{M},\Delta t}(\cdot,(n+1)\Delta t)$ are nonnegative.*

Proof. The case where $(u_K^n, v_K^n)_{K\in\mathcal{V}} \equiv 0$ has to be treated apart. In this very particular case, it is easy to check that $\left(u_K^{n+1}, v_K^{n+1}\right)_{K\in\mathcal{V}} \equiv 0$ is a solution to the scheme.

Let us now focus on the case where u_K^n or v_K^n is strictly positive for some $K \in \mathcal{V}$. Because of the weak coupling on the numerical scheme, we can first solve (4.21), and afterwards (4.13). The existence of a solution $\left(v_K^{n+1}\right)_{K\in\mathcal{V}}$ can be proved by slightly adapting the proof of [16, Proposition 3.11], which relies on a topological argument [54, 23]. The main difficulty comes from the fact the scheme (4.21) is not continuous w.r.t. $\left(v_K^{n+1}\right)_{K\in\mathcal{V}}$ on $(\mathbb{R}_+)^{\#\mathcal{V}}$, but Lemma 4.13 ensures that no component v_K^{n+1} of the discrete solution can go close to 0. Let us detail now the proof.

Let $\mu \in [0,1]$, we denote by $\left(v_{K,\mu}^{n+1}\right)_{K\in\mathcal{V}}$ the solution (if it exists) to the numerical scheme

$$\begin{aligned}\frac{v_{K,\mu}^{n+1}-v_K^n}{\Delta t}m_K+\mu\sum_{\sigma_{KL}\in\mathcal{E}_K} D_{KL}\eta_{KL,\mu}^{n+1}\left(p(v_{K,\mu}^{n+1})-p(v_{L,\mu}^{n+1})\right)\\ +(1-\mu)\sum_{\sigma_{KL}\in\mathcal{E}_K}|D_{KL}|(p(v_{K,\mu}^{n+1})-p(v_{L,\mu}^{n+1}))=\alpha u_K^n m_K-\beta v_{K,\mu}^{n+1}m_K.\end{aligned} \tag{4.50}$$

In the above scheme, we have set

$$\eta_{KL,\mu}^{n+1}=\begin{cases}\max_{v\in J_{KL,\mu}^{n+1}}\eta(v) & \text{if } D_{KL}\geq 0,\\ \min_{v\in J_{KL,\mu}^{n+1}}\eta(v) & \text{if } D_{KL}<0,\end{cases}$$

where $J_{KL,\mu}^{n+1} = \left[\min(v_{K,\mu}^{n+1}, v_{L,\mu}^{n+1}), \max(v_{K,\mu}^{n+1}, v_{L,\mu}^{n+1})\right]$. Reproducing carefully the analysis carried out in §4.5.2 and §4.5.4, we get that for all $\mu \in [0,1]$,

$$\sum_{\sigma_{KL}\in\mathcal{E}} D_{KL}\left(\phi(v_{K,\mu}^{n+1})-\phi(v_{L,\mu}^{n+1})\right)^2 \leq \sum_{\sigma_{KL}\in\mathcal{E}} D_{KL}\eta_{KL,\mu}^{n+1}\left(p(v_{K,\mu}^{n+1})-p(v_{L,\mu}^{n+1})\right)^2 \leq C \tag{4.51}$$

and, that there exists $\epsilon > 0$ such that

$$v_{K,\mu}^{n+1} \geq \epsilon > 0, \qquad \forall K \in \mathcal{V}. \tag{4.52}$$

This ensures in particular that for all $\mu \in [0,1]$, the solutions of (4.50) stay in the interior of a compact subset $\mathcal{K}$ of $\mathbb{R}^{\#\mathcal{V}}$ such that

$$\operatorname{dist}\left(\mathcal{K}, (\mathbb{R}_-)^{\#\mathcal{V}}\right) \geq \frac{\epsilon}{2}.$$

Define the function $\Upsilon : \mathcal{K}\times[0,1] \to \mathbb{R}^{\#\mathcal{V}}$ by : $\forall K \in \mathcal{V}$,

$$\begin{aligned}\Upsilon_K\left((w_K)_K,\mu\right) =& \frac{w_K-v_K^n}{\Delta t}m_K+\mu\sum_{\sigma_{KL}\in\mathcal{E}_K} D_{KL}\eta_{KL,\mu}^{n+1}\left(p(w_K)-p(w_L)\right)\\ &+(1-\mu)\sum_{\sigma_{KL}\in\mathcal{E}_K}|D_{KL}|(p(w_K)-p(w_L))-\alpha u_K^n m_K+\beta w_K m_K.\end{aligned}$$

The function Υ is uniformly continuous on $\mathcal{K}\times[0,1]$, and for all $\mu \in [0,1]$ the solution $v_{K,\mu}^{n+1}$ of the nonlinear system

$$\Upsilon\left(\left(v_{K,\mu}^{n+1}\right)_{K\in\mathcal{V}},\mu\right)=0 \tag{4.53}$$

cannot reach $\partial\mathcal{K}$. For $\mu = 0$, the system is monotone, so that the system (4.53) admits a unique solution, whose topological degree is equal to 1 (we refer to [30, Proposition 3.1] for a proof of this property). The topological degree being constant w.r.t. $\mu \in [0,1]$, the system (4.53) admits at least one solution for $\mu = 1$, concluding the proof of the existence of $\left(v_K^{n+1}\right)_{K\in\mathcal{V}}$.

The existence proof for $\left(u_K^{n+1}\right)_{K\in\mathcal{V}}$ is similar but simpler since

1. the *a priori* estimate $0 \leq u_K^{n+1} \leq 1$ is sufficient for the claim, and no energy estimate is needed here ;
2. the scheme (4.13) depends in a uniformly continuous way on $\left(u_K^{n+1}\right)_{K\in\mathcal{V}}$ on the compact subset $[-1,2]^{\#\mathcal{V}}$ of $\mathbb{R}^{\#\mathcal{V}}$.

Therefore, we let to the reader the care of checking the proof for self-conviction. □

4.6 Compactness estimates on the family of discrete solutions.

In this section, we derive estimates on differences of time and space translates of the discrete solutions necessary to prove the relative compactness property of the sequence of approximate solutions. To do this, we use the increasing function $\xi : [0,1] \longrightarrow \mathbb{R}$ and $\phi : [0,+\infty) \longrightarrow \mathbb{R}$ defined in equation (4.5).

For all $K \in \mathcal{V}_h$ and for all $n \geq 1$, we denote by $\xi_K^n = \xi\left(u_K^n\right)$, and by $\xi_{\mathcal{T}_h,\Delta t_h}$ the corresponding piecewise affine in space and constant in time reconstruction in $\mathcal{H}_{\mathcal{T}_h,\Delta t_h}$, and by $\xi_{\mathcal{M}_h,\Delta t_h}$ the piecewise constant reconstruction in $\mathcal{X}_{\mathcal{M}_h,\Delta t_h}$.

4.6.1 Time translate estimate.

We give below the time translate estimate for the family $\left(\xi_{\mathcal{M}_h,\Delta t_h}\right)_h$ expressed using the function ξ defined in (4.5). We denote by $Q_{t_\mathrm{f}-\tau} = \Omega \times (0, t_\mathrm{f}-\tau)$, for all $\tau \in (0, t_\mathrm{f})$.

Lemma 4.15. *There exists a constant $C_{\xi,\mathrm{t}}$ and $C_{v,\mathrm{t}}$ independent of h and τ such that,*

$$\iint_{Q_{t_\mathrm{f}-\tau}} \left|\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t+\tau) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)\right|^2 \mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C_{\xi,\mathrm{t}}\left(\tau + \Delta t_h\right), \tag{4.54}$$

$$\iint_{Q_{t_\mathrm{f}-\tau}} \left|v_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t+\tau) - v_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)\right|^2 \mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C_{v,\mathrm{t}}\left(\tau + \Delta t_h\right), \tag{4.55}$$

for all $\tau \in (0, t_\mathrm{f})$.

Proof. We consider the quantity $A_h(t)$ defined by

$$A_h(t) = \int_\Omega \left|\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t+\tau) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)\right|^2 \mathrm{d}\mathbf{x}, \quad \text{for all } t \in (0, t_\mathrm{f}-\tau),$$

which implies, that

$$\iint_{Q_{t_\mathrm{f}-\tau}} \left|\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t+\tau) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x},t)\right|^2 \mathrm{d}\mathbf{x}\,\mathrm{d}t = \int_0^{t_\mathrm{f}-\tau} A_h(t)\,\mathrm{d}t.$$

For $t \in (0, t_\mathrm{f}]$, we denote by $\nu(t) \in \{0,\ldots,N\}$ the unique positive integer such that $t_{\nu(t)} < t \leq t_{\nu(t+\tau)}$, so that, we can rewrite $A_h(t)$ as

$$A_h(t) = \sum_{K\in\mathcal{V}_h} \left(\xi_K^{\nu(t+\tau)+1} - \xi_K^{\nu(t)+1}\right)^2 m_K, \quad \text{for all } t \in (0, t_\mathrm{f}-\tau).$$

We observe, using the definition (4.5) of the function ξ, that

$$\begin{aligned}\left(\xi_K^{\nu(t+\tau)+1} - \xi_K^{\nu(t)+1}\right)^2 &\leq C\left(u_K^{\nu(t+\tau)+1} - u_K^{\nu(t)+1}\right) \times \left(\xi(u_K^{\nu(t+\tau)+1}) - \xi(u_K^{\nu(t)+1})\right) \\ &= C \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \left(u_K^{n+1} - u_K^n\right)\left(\xi(u_K^{\nu(t+\tau)+1}) - \xi(u_K^{\nu(t)+1})\right).\end{aligned}$$

Now, using scheme (4.13) then gathering by edges and using the weighted Young inequality, we obtain

$$\begin{aligned} A_h(t) \leq C \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} \Big[& \left(\Lambda_{KL} a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right) - \Lambda_{KL}\mu_{KL}^{n+1} a_{KL}^{n+1}\left(v_K^{n+1}-v_L^{n+1}\right)\right) \\ & \times \left(\left(\xi(u_K^{\nu(t)+1}) - \xi(u_L^{\nu(t)+1})\right) - \left(\xi(u_K^{\nu(t+\tau)+1}) - \xi(u_L^{\nu(t+\tau)+1})\right)\right) \Big] \\ & + C \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{K\in\mathcal{V}_h} m_K f\left(u_K^{n+1}\right)\left(\xi(u_K^{\nu(t+\tau)+1}) - \xi(u_K^{\nu(t)+1})\right) \\ & \leq C\left(A_{1,h}(t) + A_{2,h}(t) + A_{3,h}(t) + A_{4,h}(t) + A_{5,h}(t)\right), \end{aligned}$$

where, we have set

$$A_{1,h}(t) = \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} |\Lambda_{KL}|\, a_{KL}^{n+1}\left(u_K^{n+1}-u_L^{n+1}\right)^2,$$

$$A_{2,h}(t) = \frac{\|a\|_\infty^2 + \|\chi\|_\infty^2}{2} \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} |\Lambda_{KL}| \left(\xi(u_K^{\nu(t)+1}) - \xi(u_L^{\nu(t)+1})\right)^2,$$

$$A_{3,h}(t) = \frac{\|a\|_\infty^2 + \|\chi\|_\infty^2}{2} \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} |\Lambda_{KL}| \left(\xi(u_K^{\nu(t+\tau)+1} - \xi(u_L^{\nu(t+\tau)+1})\right)^2,$$

$$A_{4,h}(t) = \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} |\Lambda_{KL}| \left(v_K^{n+1}-v_L^{n+1}\right)^2,$$

$$A_{5,h}(t) = \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \sum_{K\in\mathcal{V}_h} m_K \left|f\left(u_K^{n+1}\right)\right| \left|\xi(u_K^{\nu(t+\tau)+1}) - \xi(u_K^{\nu(t)+1})\right|.$$

Now, we introduce the characteristic function $\rho(n,t,\tau) = 1$ if $t < n\Delta t_h \leq t+\tau$ and $\rho(n,t,\tau) = 0$ otherwise. Let $(a^n)_{n\in\{0,\dots,N\}}$ be a family of non negative real values, we have the following properties

$$\int_0^{t_f-\tau} \rho(n,t,\tau)\,\mathrm{d}t = \int_{t_n-\tau}^{t_n} \mathrm{d}t = \tau, \qquad \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h = \sum_{n;\, t<t_n<t+\tau} t_{n+1} - t_n \leq \tau + \Delta t_h,$$

and

$$\int_0^{t_f-\tau} \sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h a^{n+1}\,\mathrm{d}t = \int_0^{t_f-\tau} \sum_{n=0}^{N} \Delta t_h a^{n+1}\rho(n,t,\tau)\,\mathrm{d}t = \tau \sum_{n=0}^{N} \Delta t_h a^{n+1}.$$

Using these properties, Lemma 4.5, the a priori estimates (4.40)–(4.41) on u_K^{n+1} and v_K^{n+1}, and the L^∞ bound of the function ξ, one can deduce that

$$\int_0^{t_f-\tau} A_{1,h}(t)\,\mathrm{d}t \leq C(\tau+\Delta t_h), \quad \int_0^{t_f-\tau} A_{4,h}(t)\,\mathrm{d}t \leq C(\tau+\Delta t_h),$$

and

$$\int_0^{t_f-\tau} A_{5,h}(t)\,\mathrm{d}t \leq C(\tau+\Delta t_h),$$

for some constant $C > 0$.
It remains to show that $\int_0^{t_f-\tau} A_{2,h}(t) \leq C(\tau + \Delta t_h)$, and $\int_0^{t_f-\tau} A_{3,h}(t) \leq C(\tau + \Delta t_h)$, for some constant $C > 0$.
Consider the function ζ defined by $\zeta(n,t) = 1$ if $\nu(t) = n$ and $\zeta(n,t) = 0$ otherwise. We have the following result (see ...)

$$\int_0^{t_f-\tau} \left(\sum_{n=\nu(t)+1}^{\nu(t+\tau)} \Delta t_h \right) a^{\nu(t+\tau)+1} \, \mathrm{d}t \leq (\tau + \Delta t_h) \int_0^{t_f-\tau} \sum_{m=0}^{N} a^{m+1} \zeta(m, t+\tau) \, \mathrm{d}t$$
$$= (\tau + \Delta t_h) \sum_{m=0}^{N} a^{m+1} \int_{t^m-\tau}^{t^{m+1}-\tau} \mathrm{d}t = (\tau + \Delta t_h) \sum_{m=0}^{N} a^{m+1} \Delta t_h.$$

One can conclude the proof using this property, Lemma 4.3, and Lemma 4.5. The proof of estimate (4.55) is similar. □

We now extend by zero the functions $\xi_{\mathcal{M}_h,\Delta t_h}$ and $\xi_{\mathcal{T}_h,\Delta t_h}$ outside of Q_{t_f} and give the time translate estimate over $\mathbb{R}^3$ for $\xi_{\mathcal{M}_h,\Delta t_h}$. Indeed, there exists a constant $C > 0$ independent of h and τ such that

$$\int_{\mathbb{R}} \int_{\mathbb{R}^2} |\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t+\tau) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)|^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \leq C(\tau + \Delta t_h), \quad \text{for all } \tau \in (0, t_f).$$

Proof. Using the extension by zero of $\xi_{\mathcal{M}_h,\Delta t_h}$ outside of Q_{t_f}, one has

$$\int_{\mathbb{R}} \int_{\mathbb{R}^2} |\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t+\tau) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)|^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}t$$
$$= \iint_{Q_{t_f-\tau}} |\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t+\tau) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)|^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}t + \int_{t_f-\tau}^{t_f} \int_{\Omega} |\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)|^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}t.$$

One can deduce the proof using Lemma 4.15 and the L^∞ bound of the function ξ. □

We give now the space translate estimate on $\xi_{\mathcal{M}_h,\Delta t_h}$.

4.6.2 Space translate estimate.

Lemma 4.16. *There exists a constant $C_{\xi,s}$ independent of h and $\mathbf{y}$ such that,*

$$\int_0^{t_f} \int_{\mathbb{R}^2} |\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}+\mathbf{y}, t) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \leq C_{\xi,s}(|\mathbf{y}| + h), \tag{4.56}$$
$$\int_0^{t_f} \int_{\mathbb{R}^2} |v_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}+\mathbf{y}, t) - v_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \leq C_{\xi,s}(|\mathbf{y}| + h), \tag{4.57}$$

for all $\mathbf{y} \in \mathbb{R}^2$.

Proof. We follow the same proof used in chapter 3. We first prove that

$$\int_0^{t_f} \int_{\mathbb{R}^2} |\xi_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}+\mathbf{y}, t) - \xi_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t)| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \leq C|\mathbf{y}|, \quad \text{for all } \mathbf{y} \in \mathbb{R}^2,$$

then, using the extension by zero of $\xi_{\mathcal{M}_h,\Delta t_h}$ outside of Q_{t_f} and the triangle inequality, we get

$$\int_0^{t_f} \int_{\mathbb{R}^2} |\xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}+\mathbf{y}, t) - \xi_{\mathcal{M}_h,\Delta t_h}(\mathbf{x}, t)| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \leq A + B,$$

where

$$A = \int_0^{t_{\mathrm{f}}} \int_{\mathbb{R}^2} \left| \xi_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x}+\mathbf{y},t\right) - \xi_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t\right) \right| \mathrm{d}\mathbf{x}\,\mathrm{d}t,$$
$$B = 2 \iint_{Q_{t_{\mathrm{f}}}} \left| \xi_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t\right) - \xi_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x},t\right) \right| \mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

One can conclude the proof using the following estimate

$$\iint_{Q_{t_{\mathrm{f}}}} \left| \xi_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t\right) - \xi_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x},t\right) \right| \mathrm{d}\mathbf{x}\,\mathrm{d}t \leq Ch. \tag{4.58}$$

The proof of estimate (4.57) follows the same lines. This ends the proof of the lemma. □

4.7 Convergence

Lemma 4.17 (Strong convergence in $L^1\left(Q_{t_{\mathrm{f}}}\right)$)**.** *There exists a subsequence of the sequence* $\left(\xi\left(u\right)_{\mathcal{M}_h,\Delta t_h}\right)_{h>0}$ *(resp.* $\left(v_{\mathcal{M}_h,\Delta t_h}\right)_{h>0}$*) which converges in* $L^1\left(Q_{t_{\mathrm{f}}}\right)$ *to some function* $\xi\left(u\right) \in L^2\left(0,t_{\mathrm{f}};H^1\left(\Omega\right)\right)$ *(resp.* $v \in L^2\left(0,t_{\mathrm{f}};H^1\left(\Omega\right)\right)$*).*

Proof. Let us consider the function $\xi_{\mathcal{M}_h,\Delta t_h}$ on $Q_{t_{\mathrm{f}}}$ and its extension by zero outside of $Q_{t_{\mathrm{f}}}$. Lemma 4.15, Lemma 4.16, and the boundedness of ξ due to Lemma 4.5.1 ensure that the sequence $\left(\xi\left(u\right)_{\mathcal{M}_h,\Delta t_h}\right)_{h>0}$ verifies the assumptions of the Kolmogorov compactness criterion (see e.g. [38, 32]), where the third item of the theorem is satisfied using the triangle inequality : for any $\eta \in \mathbb{R}^2$ and $\tau \in \mathbb{R}$,

$$\begin{aligned}\left\|\xi_{\mathcal{M}_h,\Delta t_h}\left(\cdot+\eta,\cdot+\tau\right) - \xi_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)} &\leq \left\|\xi_{\mathcal{M}_h,\Delta t_h}\left(\cdot+\eta,\cdot\right) - \xi_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)} \\ &\quad + \left\|\xi_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot+\tau\right) - \xi_{\mathcal{M}_h,\Delta t_h}\left(\cdot,\cdot\right)\right\|_{L^1(\mathbb{R}^3)}.\end{aligned}$$

Kolmogorov's theorem ensures that the sequence $\left(\xi\left(u\right)_{\mathcal{M}_h,\Delta t_h}\right)_{h>0}$ is relatively compact in $L^1\left(Q_{t_{\mathrm{f}}}\right)$, that implies the existence of a subsequence of $\left(\xi\left(u\right)_{\mathcal{M}_h,\Delta t_h}\right)_{h>0}$ such that

$$\xi\left(u\right)_{\mathcal{M}_h,\Delta t_h} \longrightarrow \xi^\star \quad \text{strongly in } L^1\left(Q_{t_{\mathrm{f}}}\right). \tag{4.59}$$

Furthermore, as ξ is a continuous and nondecreasing function on $[0,1]$, there exists a unique $u\left(\mathbf{x},t\right)$ defined by

$$u\left(\mathbf{x},t\right) = \xi^{-1}\left(\xi^\star\left(\mathbf{x},t\right)\right), \qquad \text{for a.e. } \left(\mathbf{x},t\right) \in Q_{t_{\mathrm{f}}-\tau}.$$

Since ξ^{-1} is well defined and continuous, applying the L^∞ bound on $u_{\mathcal{M}_h,\Delta t_h}$ and the dominated convergence theorem to $u_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t\right) = \xi^{-1}\left(\xi_{\mathcal{M}_h,\Delta t_h}\left(\mathbf{x},t\right)\right), \quad \forall\left(\mathbf{x},t\right) \in \omega_K \times \left(0,t_{\mathrm{f}}\right), \forall K \in \mathcal{V}$, we get

$$u_{\mathcal{M}_h,\Delta t_h} \longrightarrow u \text{ a.e. in } Q_{t_{\mathrm{f}}} \text{ and strongly in } L^p\left(Q_{t_{\mathrm{f}}}\right) \text{ for } p < +\infty.$$

It remains to show that $\xi\left(u\right) \in L^2\left(0,t_{\mathrm{f}};H^1\left(\Omega\right)\right)$. Indeed, plugging estimate (4.41) into estimate (4.22), one gets that $\nabla\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}$ is uniformly bounded in $\left(L^2\left(Q_{t_{\mathrm{f}}}\right)\right)^2$. It follows that the sequence $\left(\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}\right)_h$ is bounded in $L^2\left(0,t_{\mathrm{f}};H^1\left(\Omega\right)\right)$ since $\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}$ is uniformly bounded in $L^2\left(Q_{t_{\mathrm{f}}}\right)$. Consequently, the sequence $\left(\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}\right)_h$ converges weakly, up to an unlabeled subsequence, to a function $\tilde{\xi}$ in $L^2\left(0,t_{\mathrm{f}};H^1\left(\Omega\right)\right)$.
According to estimate (4.58), the sequences $\left(\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}\right)_h$ and $\left(\xi\left(u\right)_{\mathcal{M}_h,\Delta t_h}\right)_h$ have the same limit, as a consequence $\xi\left(u\right) = \xi^\star = \tilde{\xi} \in L^2\left(0,t_{\mathrm{f}};H^1\left(Q_{t_{\mathrm{f}}}\right)\right)$. □

4.7.1 Identification as a weak solution

It remains to be shown that (u, v) is a weak solution to the continuous problem (4.1)–(4.3) in the sense of Definition 4.1. To do this, we consider a test function $\psi \in \mathcal{D}\left(\overline{\Omega} \times [0, t_{\mathrm{f}})\right)$, and denote by $\psi_K^n = \psi(\mathbf{x}_K, t_n)$, for all $K \in \mathcal{V}_h$ and all $n \in \{0, \dots, N_h\}$. Let us focus on the convergence of the first equation of scheme (4.13)–(4.21), the convergence of the second equation being similar.

Multiplying the first equation (4.13) by $\Delta t_h \psi_K^n$ and summing over $n \in \{0, \dots, N_h\}$ and $K \in \mathcal{V}_h$ yields, after a reorganization of the sum (see e.g. [34]),

$$\mathcal{A}_h + \mathcal{B}_h + \mathcal{C}_h + \mathcal{D}_h = \mathcal{F}_h, \tag{4.60}$$

where

$$\mathcal{A}_h = \sum_{n=0}^{N_h} \sum_{K \in \mathcal{V}_h} \left(u_K^{n+1} - u_K^n\right) \psi_K^n m_K, \qquad \mathcal{F}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{K \in \mathcal{V}_h} f\left(u_K^{n+1}\right) \psi_K^n m_K,$$

$$\mathcal{B}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{\sigma_{KL} \in \mathcal{E}_h} \Lambda_{KL} \left(a_{KL}^{n+1} \left(u_K^{n+1} - u_L^{n+1}\right) - \sqrt{a_{KL}^{n+1}} \left(\xi\left(u_K^{n+1}\right) - \xi\left(u_L^{n+1}\right)\right)\right) \left(\psi_K^n - \psi_L^n\right),$$

$$\mathcal{C}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{\sigma_{KL} \in \mathcal{E}_h} \Lambda_{KL} \sqrt{a_{KL}^{n+1}} \left(\xi\left(u_K^{n+1}\right) - \xi\left(u_L^{n+1}\right)\right) \left(\psi_K^n - \psi_L^n\right),$$

$$\mathcal{D}_h = -\sum_{n=0}^{N_h} \Delta t_h \sum_{\sigma_{KL} \in \mathcal{E}_h} \Lambda_{KL} \mu_{KL}^{n+1} a_{KL}^{n+1} \left(v_K^{n+1} - v_L^{n+1}\right) \left(\psi_K^n - \psi_L^n\right).$$

Time evolution term Note that $\psi_K^{N_h+1} = 0$ for all $K \in \mathcal{V}_h$, then, performing summation by parts in time, the term $\mathcal{A}_h$ can be rewritten

$$\begin{aligned}
\mathcal{A}_h &= \sum_{n=0}^{N_h} \sum_{K \in \mathcal{V}_h} u_K^{n+1} \psi_K^n m_K - \sum_{n=1}^{N_h} \sum_{K \in \mathcal{V}_h} u_K^n \psi_K^n m_K - \sum_{K \in \mathcal{V}_h} u_K^0 \psi_K^0 m_K \\
&= -\sum_{n=0}^{N_h} \Delta t_h \sum_{K \in \mathcal{V}_h} u_K^{n+1} \frac{\psi_K^{n+1} - \psi_K^n}{\Delta t_h} m_K - \sum_{K \in \mathcal{V}_h} u_K^0 \psi_K^0 m_K \\
&= -\iint_{Q_{t_{\mathrm{f}}}} u_{\mathcal{M}_h, \Delta t_h}(\mathbf{x}, t)\, \partial_t \psi_{\mathcal{M}_h, \Delta t_h}(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_\Omega u_{\mathcal{M}_h, \Delta t_h}(\mathbf{x}, 0)\, \psi_{\mathcal{M}_h, \Delta t_h}(\mathbf{x}, 0)\, \mathrm{d}\mathbf{x}.
\end{aligned}$$

Thanks to the regularity of ψ, and the uniform convergence in $L^q(Q_{t_{\mathrm{f}}})$, for all $q \in [1, \infty)$, of the sequence $(u_{\mathcal{M}_h, \Delta t_h})_h$ towards u, it follows that

$$\mathcal{A}_h \longrightarrow -\iint_{Q_{t_{\mathrm{f}}}} u(\mathbf{x}, t)\, \partial_t \psi(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_\Omega u(\mathbf{x}, 0)\, \psi(\mathbf{x}, 0)\, \mathrm{d}\mathbf{x}, \qquad \text{as } h \to 0.$$

Diffusion term Let us first prove that $\lim_{h \to 0} \mathcal{B}_h = 0$.

For all $\sigma_{KL} \in \mathcal{E}_h$ and all $n \in \{0, \dots, N_h\}$, we denote by $\overline{a}_{KL}^{n+1}$ the quantity defined by

$$\overline{a}_{KL}^{n+1} = \begin{cases} \left(\dfrac{\xi(u_K^{n+1}) - \xi(u_L^{n+1})}{u_K^{n+1} - u_L^{n+1}} \right)^2 & \text{if } u_K^{n+1} \neq u_L^{n+1}, \\ a\left(u_K^{n+1}\right) & \text{if } u_K^{n+1} = u_L^{n+1}. \end{cases}$$

Then, the term $\mathcal{B}_h$ rewrites

$$\mathcal{B}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} \Lambda_{KL}\sqrt{a_{KL}^{n+1}}\left(\sqrt{a_{KL}^{n+1}} - \sqrt{\overline{a}_{KL}^{n+1}}\right)\left(u_K^{n+1} - u_L^{n+1}\right)\left(\psi_K^n - \psi_L^n\right).$$

Now, using the Cauchy-Schwarz inequality, we get

$$|\mathcal{B}_h| \leq \left(\sum_{n=0}^{N_h} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} |\Lambda_{KL}|\, a_{KL}^{n+1}\left(u_K^{n+1} - u_L^{n+1}\right)^2\right)^{\frac{1}{2}} \times \mathcal{R}_h{}^{\frac{1}{2}},$$

where, $\mathcal{R}_h$ is given by

$$\mathcal{R}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{\sigma_{KL}\in\mathcal{E}_h} |\Lambda_{KL}|\left(\sqrt{a_{KL}^{n+1}} - \sqrt{\overline{a}_{KL}^{n+1}}\right)^2 \left(\psi_K^n - \psi_L^n\right)^2.$$

Using Lemma 4.5 and Proposition 4.9, one has $|\mathcal{B}_h| \leq C\mathcal{R}_h{}^{\frac{1}{2}}$. However, in order to prove that $\lim_{h\to 0} \mathcal{B}_h = 0$, it suffices to prove that $\lim_{h\to 0} \mathcal{R}_h = 0$.
For all $T \in \mathcal{T}_h$, we denote by

$$\overline{\xi}_T^{n+1} = \max_{\mathbf{x}\in T}\left(\xi\left(p\right)_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t_{n+1})\right), \qquad \underline{\xi}_T^{n+1} = \min_{\mathbf{x}\in T}\left(\xi\left(p\right)_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t_{n+1})\right),$$

and for all $(\mathbf{x}, t) \in T \times (t_n, t_{n+1})$, by

$$\overline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t) = \overline{\xi}_T^{n+1}, \qquad \underline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t) = \underline{\xi}_T^{n+1}.$$

Consider the uniform continuous function $\sqrt{a\circ\xi^{-1}}$ defined on the closed and bounded interval $[0, \xi(1)]$, and let η be its modulus of continuity, then we have

$$\left|\sqrt{a_{KL}^{n+1}} - \sqrt{\overline{a}_{KL}^{n+1}}\right| \leq \eta\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right), \qquad \text{for all } \sigma_{KL} \in \mathcal{E}_T. \tag{4.61}$$

Therefore, using this inequality in the definition of $\mathcal{R}_h$, we get

$$0 \leq \mathcal{R}_h \leq \mathcal{Q}_h \tag{4.62}$$

where,

$$\mathcal{Q}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{T\in\mathcal{T}_h}\left(\eta\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} \left|\lambda_{KL}^T\right|\left(\psi_K^n - \psi_L^n\right)^2. \tag{4.63}$$

Thanks to Lemma 4.4, one can deduce that the inequality (4.62) implies that

$$0 \leq \mathcal{R}_h \leq C\iint_{Q_{t_f}} \eta\left(\overline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t) - \underline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t)\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

where C is independent of h, and Δt_h.
Therefore, it suffices to show that $\overline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t) - \underline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t) \to 0$ for a.e. in Q_{t_f} to consequently prove that $\lim_{h\to 0} \mathcal{R}_h = 0$. However, by a simple generalization of Lemma A.1 and by the help of Lemma 4.3 and Proposition 4.9, it follows that

$$\iint_{Q_{t_f}} \left|\overline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t) - \underline{\xi}_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t)\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq Ch\left(\iint_{Q_{t_f}} \left|\nabla\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}, t)\right|^2 \mathrm{d}\mathbf{x}\,\mathrm{d}t\right)^{\frac{1}{2}} \leq Ch.$$

As a consequence, up to a subsequence, one has

$$\lim_{h\to 0}\mathcal{B}_h=\lim_{h\to 0}\mathcal{R}_h=\lim_{h\to 0}\mathcal{Q}_h=0.$$

We now focus on the term $\mathcal{C}_h$ and prove that

$$\lim_{h\to 0}\mathcal{C}_h=\iint_{Q_{t_{\mathrm{f}}}}\Lambda\left(\mathbf{x}\right)a\left(u\right)\nabla u\cdot\nabla\psi\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

To do this, we introduce the term $\mathcal{C}_h'$ defined by

$$\mathcal{C}_h':=\iint_{Q_{t_{\mathrm{f}}}}\Theta_{\mathcal{T}_h,\Delta t_h}\Lambda\left(\mathbf{x}\right)\nabla\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}\cdot\nabla\psi_{\mathcal{T}_h,\Delta t_h}\left(\cdot,t-\Delta t_h\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

where $\Theta_{\mathcal{T}_h,\Delta t_h}$ is a continuous function given by

$$\Theta_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x},t\right)=\sqrt{a\circ\xi^{-1}}\left(\Upsilon_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x},t\right)\right),\qquad\forall\mathbf{x}\in T,\,\forall t\in\left(t_n,t_{n+1}\right],\,\forall T\in\mathcal{T}_h,$$

and $\Upsilon_{\mathcal{T}_h,\Delta t_h}$ is a function defined by

$$\Upsilon_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x},t\right)=\xi\left(u\right)_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x}_T,t\right),\qquad\forall\mathbf{x}\in T,\,\forall t\in\left(t_n,t_{n+1}\right],\,\forall T\in\mathcal{T}_h.$$

Using again a generalization of Lemma A.1 as well as the boundedness of the continuous function $\sqrt{a\circ\xi^{-1}}$, we obtain

$$\begin{aligned}\Upsilon_{\mathcal{T}_h,\Delta t_h}&\longrightarrow\xi\left(u\right)&&\text{in }L^2\left(Q_{t_{\mathrm{f}}}\right)\text{ as }h\to 0,\\ \Theta_{\mathcal{T}_h,\Delta t_h}&\longrightarrow\sqrt{a\left(u\right)}&&\text{in }L^2\left(Q_{t_{\mathrm{f}}}\right)\text{ as }h\to 0.\end{aligned}\tag{4.64}$$

It remains to verify that $|\mathcal{C}_h-\mathcal{C}_h'|\longrightarrow 0$, when h tends to zero.
We denote by

$$a_T^{n+1}=\left(\Theta_{\mathcal{T}_h,\Delta t_h}\left(\mathbf{x}_T,t_{n+1}\right)\right)^2,\qquad\forall T\in\mathcal{T}_h,\,\forall n\in\{0,\ldots,N_h\}.$$

The discretization of the term $\mathcal{D}_m'$ is written as

$$\mathcal{C}_m'=\sum_{n=0}^{N_h}\Delta t_h\sum_{T\in\mathcal{T}_h}\sqrt{a_T^{n+1}}\sum_{\sigma_{KL}\in\mathcal{E}_T}\lambda_{KL}^T\left(\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right)\left(\psi_K^n-\psi_L^n\right).$$

Performing the same way as for the inequality (4.61), one has

$$\left|\sqrt{a_{KL}^{n+1}}-\sqrt{a_T^{n+1}}\right|\leq\eta\left(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}\right),\qquad\text{for all }\sigma_{KL}\in\mathcal{E}_T.$$

Therefore, using the Cauchy-Schwarz inequality, Lemma 4.3, Lemma 4.4, and Proposition 4.9, we deduce that there exists a constant C does not depend on h such that

$$\begin{aligned}\left|\mathcal{C}_m-\mathcal{C}_m'\right|^2&\leq\left(\sum_{n=0}^{N_h}\Delta t_h\sum_{T\in\mathcal{T}_h}\eta\left(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}\right)\sum_{\sigma_{KL}\in\mathcal{E}_T}\left|\lambda_{KL}^T\right|\left|\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right|\left|\psi_K^n-\psi_L^n\right|\right)^2\\&\leq\mathcal{Q}_h\times\sum_{n=0}^{N_h}\Delta t_h\sum_{T\in\mathcal{T}_h}\sum_{\sigma_{KL}\in\mathcal{E}_T}\left|\lambda_{KL}^T\right|\left|\xi\left(u_K^{n+1}\right)-\xi\left(u_L^{n+1}\right)\right|^2\\&\leq C\mathcal{Q}_h\longrightarrow 0\quad\text{as }h\to 0.\end{aligned}$$

Convection term For all $T \in \mathcal{T}_h$, we define the piecewise constant function $\kappa_{\mathcal{T}_h,\Delta t_h}$ by

$$\kappa_{\mathcal{T}_h,\Delta t_h}(\mathbf{x},t) = \chi \circ \xi^{-1}\left(\Upsilon_{\mathcal{T}_h,\Delta t_h}(\mathbf{x},t)\right), \qquad \forall \mathbf{x} \in T,\ \forall t \in (t_n, t_{n+1}].$$

Using the same guidelines as for the convergence results (4.64), one has

$$\kappa_{\mathcal{T}_h,\Delta t_h} \longrightarrow \chi(u) \qquad \text{in } L^2(Q_{t_\mathrm{f}}) \text{ as } h \to 0.$$

We introduce the term

$$\mathcal{D}'_h := -\iint_{Q_{t_\mathrm{f}}} \kappa_{\mathcal{T}_h,\Delta t_h} \Lambda(\mathbf{x}) \nabla v_{\mathcal{T}_h,\Delta t_h} \cdot \nabla \psi_{\mathcal{T}_h,\Delta t_h}(\cdot, t-\Delta t_h)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Thanks to the weakly convergence in $L^2(Q_{t_\mathrm{f}})$ of the sequence $\nabla v_{\mathcal{T}_h,\Delta t_h}$ towards ∇v, and to the uniform convergence of $\nabla \psi_{\mathcal{T}_h,\Delta t_h}$ towards $\nabla \psi$, we obtain

$$\mathcal{D}'_h \longrightarrow -\iint_{Q_{t_\mathrm{f}}} \chi(u)\Lambda(\mathbf{x})\nabla v \cdot \nabla \psi \mathrm{d}\mathbf{x}\,\mathrm{d}t \qquad \text{as } h \to 0.$$

Let us prove, using the same guidelines as before, that $|\mathcal{D}_h - \mathcal{D}'_h| \longrightarrow 0$, when h tends to zero. We denote by

$$\begin{aligned} \chi_T^{n+1} &= \kappa_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}_T, t_{n+1}), \qquad \forall T \in \mathcal{T}_h, \forall n \in \{0,\dots,N_h\}, \\ \mu_T^{n+1} &= \mu_{\mathcal{T}_h,\Delta t_h}(\mathbf{x}_T, t_{n+1}), \qquad \forall T \in \mathcal{T}_h, \forall n \in \{0,\dots,N_h\}. \end{aligned}$$

Therefore,

$$\mathcal{D}_h - \mathcal{D}'_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{T\in\mathcal{T}_h} \sum_{\sigma_{KL}\in\mathcal{E}_T} \left(a_T^{n+1}\mu_T^{n+1} - a_{KL}^{n+1}\mu_{KL}^{n+1}\right) \lambda_{KL}^T \left(v_K^{n+1} - v_L^{n+1}\right)\left(\psi_K^n - \psi_L^n\right).$$

Thanks to the the triangle inequality and to the existence of a continuity moduli η and δ of the continuous functions $\sqrt{a \circ \xi^{-1}}$ and $\mu \circ \xi^{-1}$ respectively, one has

$$\begin{aligned} \left|a_{KL}^{n+1}\mu_{KL}^{n+1} - a_T^{n+1}\mu_T^{n+1}\right| &\le \mu_{KL}^{n+1}\left|a_{KL}^{n+1} - a_T^{n+1}\right| + a_T^{n+1}\mu_{KL}^{n+1} - \mu_T^{n+1} \\ &\le C\left(\eta\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right) + \delta\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)\right), \end{aligned}$$

where the constant C does not depend on h. Therefore, using the Cauchy-Schwarz inequality, Lemma 4.3, Lemma 4.4, and Proposition 4.9, we deduce that there exists a constant C independent of h such that

$$\left|\mathcal{D}_h - \mathcal{D}'_h\right|^2 \le C\left(\mathcal{Q}_h + \mathcal{W}_h\right) \times \sum_{n=0}^{N_h} \Delta t_h \sum_{T\in\mathcal{T}_h} \sum_{\sigma_{KL}\in\mathcal{E}_T} \left|\lambda_{KL}^T\right| \left|v_K^{n+1} - v_L^{n+1}\right|^2,$$

where Q_h is given by equation (4.63), and $\mathcal{W}_h$ is given by

$$\mathcal{W}_h = \sum_{n=0}^{N_h} \Delta t_h \sum_{T\in\mathcal{T}_h} \left(\delta\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} \left|\lambda_{KL}^T\right| \left(\psi_K^n - \psi_L^n\right)^2.$$

Now, using the same proof as for the diffusive term, one can deuce that $\mathcal{W}_h \le Ch$. Therefore

$$\lim_{h\to 0} \left|\mathcal{D}_h - \mathcal{D}'_h\right| = 0,$$

and consequently,

$$\lim_{h\to 0} \mathcal{D}_h = -\iint_{Q_{t_\mathrm{f}}} \chi(u)\Lambda(\mathbf{x})\nabla v \cdot \nabla \psi \mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Reaction term We would now like to show that

$$\mathcal{F}_h \longrightarrow \iint_{Q_{t_f}} f\left(u\left(\mathbf{x},t\right)\right)\psi\left(\mathbf{x},t\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \quad \text{as } h\to 0.$$

For this purpose, we denote, for all $K\in\mathcal{V}_h$ and for all $n\geq 1$, by $f_K^n=f\left(u_K^n\right)$, and by $f_{\mathcal{M}_h,\Delta t_h}$ the piecewise constant reconstruction in $\mathcal{X}_{\mathcal{M}_h,\Delta t_h}$. Thus we have

$$\mathcal{F}_h=\iint_{Q_{t_f}} f_{\mathcal{M}_h,\Delta t_h}\psi_{\mathcal{M}_h,\Delta t_h}\left(\cdot,t-\Delta t_h\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \longrightarrow \iint_{Q_{t_f}} f\left(u\left(\mathbf{x},t\right)\right)\psi\left(\mathbf{x},t\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \quad \text{as } h\to 0,$$

since $f\left(u\right)_{\mathcal{M}_h,\Delta t_h}$ converges strongly in $L^2\left(Q_{t_f}\right)$ towards $f\left(u\right)$, and as $\psi_{\mathcal{M}_h,\Delta t_h}$ converges uniformly towards ψ. This ends the proof of the convergence.

4.8 Numerical results

In this section, we establish various 2–D numerical results provided by the *nonlinear CVFE scheme* (4.13)–(4.21). Newton's algorithm is carried out for the implementation of the scheme, coupled with a biconjugate gradient method to solve linear systems arising from the Newton algorithm. We provide three tests to show the effectiveness of the *nonlinear CVFE scheme* (4.13)–(4.21). For these tests, we consider the following data : $L_{\mathbf{x}}=1$, $L_{\mathbf{y}}=1$ (the length and the width of the domain). We fix : $\Delta t=0.002$, $\alpha=0.01$, $\beta=0.05$, $a\left(u\right)=d_u u\left(1-u\right)$, $d_u=0.0005$, $\chi\left(u\right)=\zeta\times\left(u\left(1-u\right)\right)^2$, $\zeta=0.05$. By definition, we have $\mu\left(u\right)=\frac{\zeta}{d_u}u\left(1-u\right)$ then, the numerical flux function μ_{KL}^{n+1} is given using the following functions :

$$\mu_{\uparrow}\left(z\right)=\mu\left(\min\{z,\frac{1}{2}\}\right),\quad \text{and } \mu_{\downarrow}\left(z\right)=\mu\left(\max\{z,\frac{1}{2}\}\right)-\mu\left(\frac{1}{2}\right),\quad \forall z\in\left(0,1\right)\times\left(0,1\right).$$

Unless stated otherwise and throughout the tests, we assume that $f\left(u\right)=0$, that the initial conditions are defined by regions, and we assume zeros-flux boundary conditions. For instance, the cell density is initially defined by $u_0\left(\mathbf{x},\mathbf{y}\right)=1$ in the square region given by $\left(\mathbf{x},\mathbf{y}\right)\in[0.45,\,0.55]$ and 0 otherwise. The initial chemeoattractant concentration is defined by $v_0\left(\mathbf{x},\mathbf{y}\right)=5$ in the space region given by $\left(\mathbf{x},\mathbf{y}\right)\in[0.2,\,0.3]\times[0.45,\,0.55]\cup[0.45,\,0.55]\times[0.2,\,0.3]\cup[0.45,\,0.55]\times[0.7,\,0.8]\cup[0.7,\,0.8]\times[0.45,\,0.55]$.

Test 1 (Weak anisotropic case). In this test, we assume that the diffusion tensors are given by

$$\Lambda\left(\mathbf{x}\right)=\begin{pmatrix}1 & 0\\ 0 & \theta\end{pmatrix},\quad D\left(\mathbf{x}\right)=dI_2,\quad d=0.0001.$$

Further, we consider an admissible triangular primary mesh made of 14 336 triangles, the corresponding Donald dual mesh consists of 7 297 dual control volumes. In a admissible triangular mesh, all the angles of triangles are acute, then one can deduce that the transmissibility coefficients are nonnegative in the case of isotropic diffusion tensors. In Tab. 4.1, we present minimum and maximum values obtained with each of the scheme (3.24)–(3.25) (of the previous chapter), the *nonlinear CVFE scheme* (4.13)–(4.21), and the FV scheme.

Test 2 (Weak anisotropic case/obtuse angles). In this test, we consider a general unstructured mesh that contains obtuse angles, this mesh is made of 5 193 triangles and 2 665 dual control volumes (see Figure 4.2 for an illustration of the primal and dual mesh).

		scheme (3.24)–(3.25)	*scheme* (4.13)–(4.21)	FV scheme
After 1 iteration $\theta = 1$	Min. Val. u Max. Val. u	0.000000 1.000000	0.000000 1.000000	0.000000 1.000000
After 10 iterations $\theta = 1$	Min. Val. u Max. Val. u	0.000000 0.971110	0.000000 1.000000	0.000000 1.000000
After 1 iteration $\theta = 5$	Min. Val. u Max. Val. u	-1.73001×10^{-3} 0.99722922	8.68789×10^{-20} 1.000000	| |
After 10 iterations $\theta = 5$	Min. Val. u Max. Val. u	-1.62500×10^{-2} 0.9715705	0.000000 1.000000	| |
After 1 iteration $\theta = 10$	Min. Val. u Max. Val. u	-4.46953×10^{-3} 1.00018368	6.30555×10^{-16} 1.000000	| |
After 10 iterations $\theta = 10$	Min. Val. u Max. Val. u	-3.91245×10^{-2} 0.98342428	6.30554×10^{-16} 0.9999999	| |

TABLE 4.1 – Numerical results after 1 and 10 iterations.

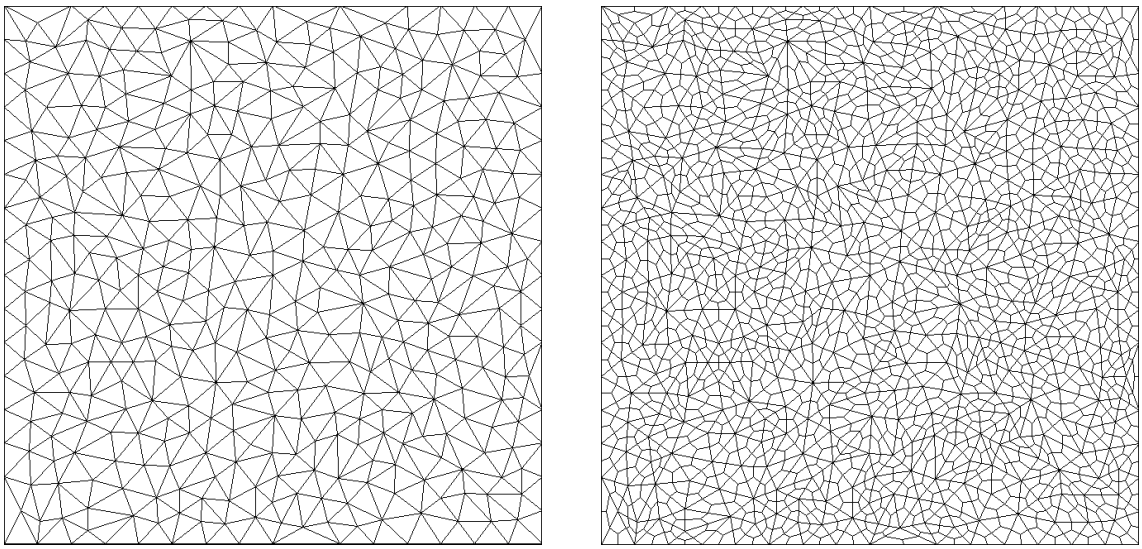


FIGURE 4.2 – Initial primal mesh $\mathcal{T}_h$ and barycentric dual mesh $\mathcal{M}_h$.

The diffusion tensors are defined, for all $\mathbf{x} \in (0,1) \times (0,1)$, by

$$\Lambda(\mathbf{x}) = \begin{pmatrix} 7 & 2 \\ 2 & 10 \end{pmatrix}, \quad D(\mathbf{x}) = dI_2, \quad d = 0.0001.$$

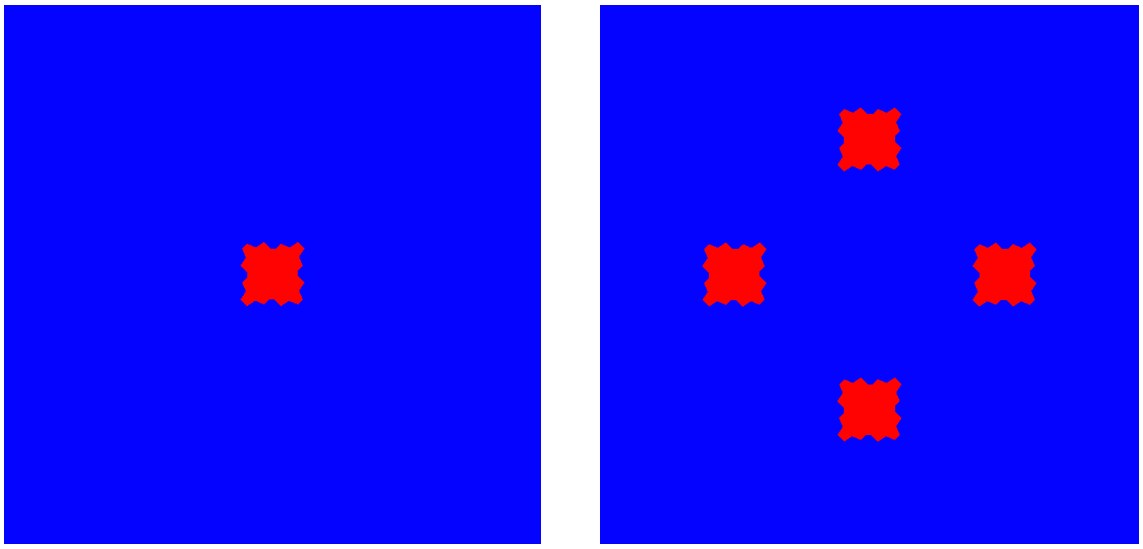


FIGURE 4.3 – Initial condition for the cell density u (left) with $0 \leq u \leq 1$ and for the chemeoattractant concentration v (right) with $0 \leq v \leq 5$.

The cell density diffusion tensor Λ is taken to be a homogeneous anisotropic tensor with high diffusivity in a direction at 63.44 degrees form the horizontal and low diffusivity in the orthogonal direction. Indeed, we have

$$\Lambda = R_\theta \times \begin{pmatrix} 11 & 0 \\ 0 & 6 \end{pmatrix} \times R_\theta^{-1},$$

where, $R_\theta = \begin{pmatrix} 0.4472 & -0.8944 \\ 0.8944 & 0.4472 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ is the rotation of angle $\theta = 63.44$ degrees.

Figure 4.3 represents initial distributions of the cell density u and the chemeoattractant concentration v over the initial triangular mesh as well as the corresponding dual mesh.

Figures 4.4–4.5 represent the evolution of the cell density at time $t = 0.4$, $t = 1.4$, $t = 2.4$, and $t = 4$. At moment $t = 0.4$, it is clear that the cell density diffuses in the space (in a direction at 63.44 degrees from the horizontal) without any interactions with the chemeoattractant which diffuses uniformly in the space. Then, after a while, and when the chemeoattractant diffusion reaches the cell density location, we see that the latter changes its direction to be absorbed by the chemeoattractant located vertically. This process continues and the cells accumulate into the location of the chemeoattractant and finally we obtain the cell density aggregations as shown at $t = 4$.

Test 3 (Anisotropic case/obtuse angles). In this test, we consider an unstructured mesh consisting of 15 568 primal triangles and 7 912 dual control dual volumes. Further, we assume that the diffusion tensors are anisotropic and are given by :

$$\Lambda(\mathbf{x}) = \begin{pmatrix} 8 & -7 \\ -7 & 20 \end{pmatrix}, \quad D(\mathbf{x}) = d \begin{pmatrix} 1 & 0 \\ 0 & \theta \end{pmatrix}, \quad d = 0.0001.$$

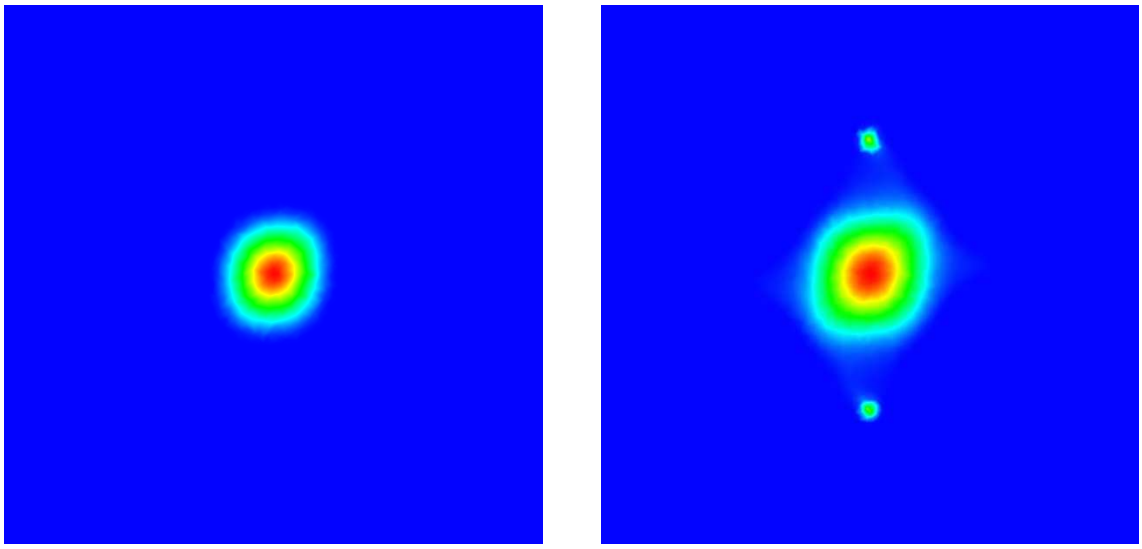


FIGURE 4.4 – Evolution of the cell density u at time $t = 0.4$ with $0 \leq u \leq 0.667$ (left), and at time $t = 1.4$ with $0 \leq u \leq 0.632$(right).

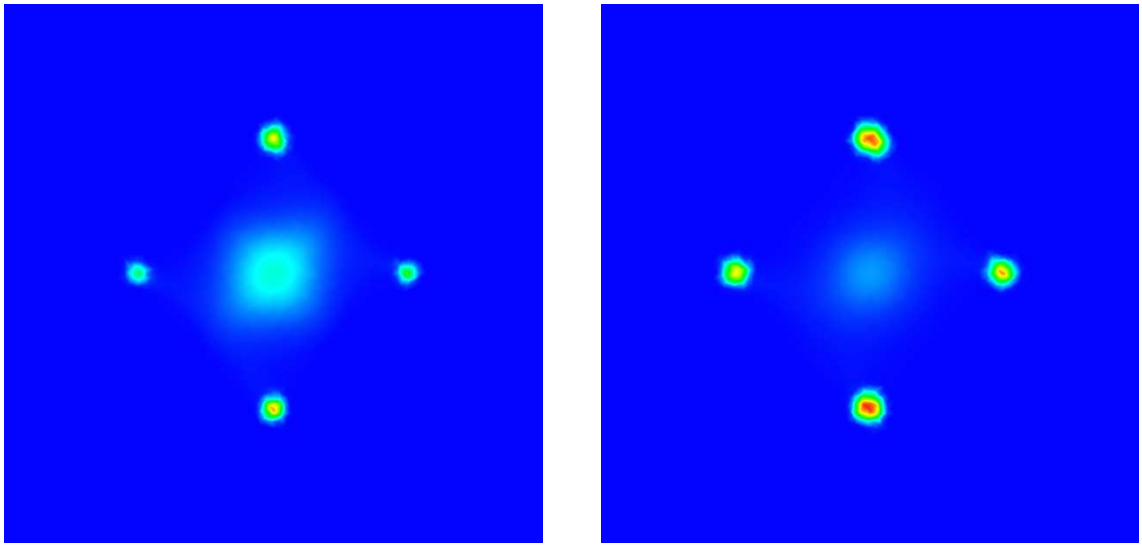


FIGURE 4.5 – Evolution of the cell density u at time $t = 2.4$ with $0 \leq u \leq 0.972$ (left), and at time $t = 4$ with $0 \leq u \leq 0.987$(right).

Table 4.2 provides a comparison between the *nonlinear CVFE scheme* coupled on the one hand with the discretization (4.15) of v and with the discretization (4.21) of v on the other hand. We see that the discretization (4.21) carries out a better approximation than the discretization (4.15) in terms of ensuring the discrete maximum principle property.

		scheme (4.13)–(4.15)	*scheme* (4.13)–(4.21)
After 1 iteration $\theta = 3$	Min. Val. u	0.0000000	0.0000000
	Max. Val. u	1.0000000	1.0000000
	Min. Val. v	-1.141912E-002	1.922764E-051
	Max. Val. v	5.012383	4.999982
After 200 iterations $\theta = 3$	Min. Val. u	0.0000000	0.0000000
	Max. Val. u	0.5298226	0.5312562
	Min. Val. v	-1.731068E-003	1.297192E-080
	Max. Val. v	4.8053827	4.8018742
After 1000 iterations $\theta = 3$	Min. Val. u	0.0000000	0.0000000
	Max. Val. u	0.9957580	0.9974757
	Min. Val. v	6.265859E-023	3.171769E-080
	Max. Val. v	2.961761	2.910828

TABLE 4.2 – Numerical results after 1, 200 and 1000 iterations over an unstructured mesh with obtuse angles.

5

Analysis of a nonlinear degenerate parabolic equation arising in chemotaxis or in porous media

Sommaire

5.1 Introduction

This chapter is devoted to the theoretical analysis of a general degenerate nonlinear parabolic equation. This kind of equations stems either from the modeling of a compressible two phase flow in porous media or from the modeling of the chemotaxis-fluid process.

In the degenerate equation, the strong nonlinearities are technically difficult to be controlled by the degenerate dissipative term because the equation itself presents degenerate terms of order 0 and of order 1.

In the case of the degeneracy of the dissipative term at one point, a weak and classical formulation is possible for the expected solutions. However, in the case of the degeneracy of the dissipative term at two points, we obtain solutions in a weaker sense compared to the one of the

classical formulation. Therefore, a degenerate weighted formulation is introduced taking into account the degeneracy of the dissipative term.

5.2 The nonlinear degenerate model

Let $T > 0$ be a fixed time and Ω be an open bounded subset of $\mathbb{R}^d$, $d = 2, 3$. We set $Q_T := \Omega \times (0, T)$ and $\Sigma_T = \partial\Omega \times (0, T)$. We consider the following nonlinear degenerate parabolic equation

$$\partial_t u - \operatorname{div}\left(a\left(u\right)\nabla u - f\left(u\right)\mathbf{V}\right) - g\left(u\right)\operatorname{div}\left(\mathbf{V}\right) + \gamma a\left(u\right)\nabla u \cdot \tilde{\mathbf{V}} = 0, \quad \text{in } Q_T. \tag{5.1}$$

The boundary condition is defined, on the saturation u, by

$$u\left(\mathbf{x}, t\right) = 0, \qquad \text{on } \Sigma_T. \tag{5.2}$$

The initial condition is given by

$$u\left(\mathbf{x}, 0\right) = u_0\left(\mathbf{x}\right), \quad \text{in } \Omega. \tag{5.3}$$

Many physical models lead to degenerate nonlinear parabolic problem. For instance, in [39] the authors analyzed a model of a degenerate nonlinear system arising from compressible two-phase flows in porous media (in this case we have $\mathbf{V} = \tilde{\mathbf{V}}$). The described system coupled the saturation (denoted by u) and the global pressure (denoted by p). The global velocity (denoted by $\mathbf{V}$) is taken to be proportional to the gradient of the global pressure. On the other hand, several papers are devoted to the mathematical analysis of compressible, miscible displacement models in porous media (see e.g. [5, 6, 35]). Here, we consider a generalization of the saturation equation where we assume that the velocity field is given and fixed.

Other models that can lead to such kind of degenerate nonlinear parabolic equation (5.1) are the chemotaxis models, where u represents the cell density and $\mathbf{V}$ represents the gradient of the chemical concentration (see e.g. [8, 26, 57]), while $\tilde{\mathbf{V}}$ represents the velocity of the fluid which transports the cell density and the chemical concentration.

We introduce the classical assumptions for porous media made about the system (5.1)–(5.3) :

(H1) The velocities $\mathbf{V}$ and $\tilde{\mathbf{V}}$ are two measurable functions lying into $\left(L^\infty\left(\Omega\right)\right)^d$. In addition, we assume the following condition on the velocity flied $\mathbf{V}$:

$$\mathbf{V} \cdot \mathbf{n} \leq 0, \quad \text{on } \Sigma_T,$$

where $\mathbf{n}$ is the unit normal vector to $\partial\Omega$ outward to Ω.

(H2) f is a differentiable function in $[0, 1]$ and $g \in \mathcal{C}^1\left([0, 1]\right)$ verifying

$$g\left(0\right) = f\left(0\right) = 0, \quad f\left(1\right) = g\left(1\right) = 1, \text{ and } g'\left(u\right) \geq C_{g'} > 0 \quad \forall u \in [0, 1].$$

(H3) The initial condition u_0 satisfies : $u_0 \in L^2\left(\Omega\right)$ and $0 \leq u_0\left(\mathbf{x}\right) \leq 1$ for a.e. $\mathbf{x} \in \Omega$.

A major difficulty of system (5.1)–(5.3) is the possible degeneracy of the diffusion term. Here, we give the degeneracy assumption on the dissipation function a :

(H4a) $a \in \mathcal{C}^1([0, 1], \mathbb{R})$, $a\left(u\right) > 0$ for $0 < u < 1$, $a\left(0\right) > 0$, $a\left(1\right) = 0$.
Furthermore, there exist $a_0 > 0$, $0 < r_2 \leq 2$, $u_* < 1$, m_1 and $M_1 > 0$ such that
$a\left(u\right) \geq a_0$ for all $0 \leq u \leq u_*$,
$m_1\left(1 - u\right)^{r_2} \leq a\left(u\right) \leq M_1\left(1 - u\right)^{r_2}$, for all $u_* \leq u \leq 1$. In addition there exists c_1, $c_2 > 0$ such that $c_1\left(1 - u\right)^{-1} \leq \left(f\left(u\right) - g\left(u\right)\right)^{-1} \leq c_2\left(1 - u\right)^{-1}$ for all $u_* \leq u < 1$.

(H4b) $a \in \mathcal{C}^1([0,1],\mathbb{R})$, $a(u) > 0$ for $0 < u < 1$, $a(0) = 0$, $a(1) > 0$,
Furthermore, there exist $r_1 > 0$, m_1 and $M_1 > 0$ such that
$m_1 r_1 u^{r_1-1} \leq a'(u) \leq M_1 r_1 u^{r_1-1}$, for all $0 \leq u \leq 1$. In addition, there exists a constant $C > 0$ such that $|f(u) - g(u)| \leq Cu$, $\forall 0 \leq u \leq 1$.

(H4c) $a \in \mathcal{C}^1([0,1],\mathbb{R})$, $a(u) > 0$ for $0 < u < 1$, $a(0) = 0$, $a(1) = 0$,
Furthermore, there exist $r_1 > 0$, $r_2 > 0$, $u_* < 1$, m_1 and $M_1 > 0$ such that
$m_1 r_1 u^{r_1-1} \leq a'(u) \leq M_1 r_1 u^{r_1-1}$, for all $0 \leq u \leq u_*$,
$-r_2 M_1 (1-u)^{r_2-1} \leq a'(u) \leq -r_2 m_1 (1-u)^{r_2-1}$, for all $u_* \leq u \leq 1$. In addition, there exists c_1, c_2, $C > 0$ such that $|f(u) - g(u)| \leq Cu$ for all $0 \leq u \leq u_*$ et $c_1 (1-u)^{-1} \leq (f(u) - g(u))^{-1} \leq c_2 (1-u)^{-1}$ for all $u_* \leq u < 1$.

In what follows, we introduce first the existence of classical weak solutions to equation (5.1) under the assumptions (H1)–(H3) and (H4a) and for a particular choice of the initial data. Next, we show the existence of solutions to equation (5.1) in a weak sense (by introducing a weighted formulation) and under the assumptions (H1)–(H3) and (H4b). Finally, for a particular choice of the initial data, we introduce also the existence of solutions to equation (5.1) (verifying a weighted formulation) and under the assumptions (H1)–(H3) and (H4c).
In the sequel and for the simplicity, we assume that the nonnegative constant γ is fixed equals to 1 and that $\tilde{\mathbf{V}} = \mathbf{V}$ (the same analysis is possible with $\tilde{\mathbf{V}} \neq \mathbf{V}$).

5.2.1 Classical weak solutions

Let us consider the function $k \in \mathcal{C}^1[0,1)$ defined by

$$\begin{aligned} k(u) &= u, && \text{if } 0 \leq u \leq u_*, \\ k'(u) &= (f(u) - g(u))^{-1} g'(u)\, k(u), && \text{if } u_* \leq u < 1. \end{aligned} \tag{5.4}$$

We denote by L the following primitive of the function k

$$L(u) = \int_0^u k(\tau)\, \mathrm{d}\tau, \qquad \forall\, 0 \leq u < 1.$$

Definition 5.1. Under the assumptions (H1)–(H3) and (H4a), and assume that the initial condition $u(t=0) = u_0$ satisfies $L(u_0) \in L^1(\Omega)$. We say that u is a classical weak solution to system (5.1)–(5.3) if u verifies

$$\begin{aligned} &0 \leq u(\mathbf{x},t) \leq 1 \text{ for a.e.} (\mathbf{x},t) \in \Omega \times (0,T), \\ &u \in L^2\left(0,T; H_0^1(\Omega)\right) \cap \mathcal{C}^0\left([0,T]; L^2(\Omega)\right), \\ &\partial_t u \in L^2\left(0,T; H^{-1}(\Omega)\right), \end{aligned}$$

and such that

$$\begin{aligned} \int_0^T \langle \partial_t u, \varphi \rangle\, \mathrm{d}t &+ \int_{Q_T} a(u) \nabla u \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_{Q_T} f(u) \mathbf{V} \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &+ \int_{Q_T} g(u) \mathbf{V} \cdot \nabla \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t + \int_{Q_T} g'(u) \mathbf{V} \cdot \nabla u \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &+ \int_{Q_T} a(u) \mathbf{V} \cdot \nabla u \varphi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t = 0, \qquad \forall \varphi \in L^2\left(0,T; H_0^1(\Omega)\right). \end{aligned} \tag{5.5}$$

where the bracket $\langle \cdot, \cdot \rangle$ represents the duality product between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Theorem 5.2 (Degenerate system)**.** *Under the assumptions* (H1) − (H3) *and* (H4a)*, there exists at least one classical weak solution to the degenerate system* (5.1)–(5.3) *in the sense of Definition 5.1.*

5.2.2 Weak degenerate solutions

Let us introduce the functions β and h defined by

$$\beta(u) = u^{r-1}, \quad h(u) = \int_0^u \beta(\tau)\,\mathrm{d}\tau, \text{ where } r = \begin{cases} r_1 + 2, & \text{if } r_1 \leq 1, \\ r_1, & \text{if } r_1 > 1. \end{cases}$$

Here, r_1 is the same constant defined in assumption (H4b).
Further, we consider the functions β_θ and h_θ defined, for all $\theta > 0$, by

$$\beta_\theta(u) = u^{r-1+\theta}, \quad h_\theta(u) = \int_0^u \beta_\theta(\tau)\,\mathrm{d}\tau.$$

Definition 5.3. For $\theta \geq 7r_1 + 6 - r$, and under the assumptions (H1)–(H3) and (H4b). We say that u is a degenerate weak solution to system (5.1)–(5.3) if

$$\begin{aligned} &0 \leq u(\mathbf{x},t) \leq 1 \text{ for a.e. } (\mathbf{x},t) \in \Omega \times (0,T), \\ &h_\theta(u) \in L^2\big(0,T; H_0^1(\Omega)\big), \\ &\sqrt{\beta'(u)}\nabla u \in \big(L^2(Q_T)\big)^d, \end{aligned}$$

and such that the function F defined, for all $\chi \in L^2\left(0,T; H^1(\Omega)\right)$ by

$$\begin{aligned} F(u,\chi) = &-\int_{Q_T} h_\theta(u)\,\partial_t\chi \mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_\Omega h_\theta(u_0)\,\chi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} + \int_{Q_T} a(u)\,\nabla u \cdot \nabla\left(\beta_\theta(u)\,\chi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \int_{Q_T} a(u)\,\mathbf{V}\cdot\nabla u\beta_\theta(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_T}\left(f(u) - g(u)\right)\mathbf{V}\cdot\nabla\left(\beta_\theta(u)\,\chi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \int_{Q_T} g'(u)\,\mathbf{V}\cdot\nabla u\beta_\theta(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_T}\left(f(u) - g(u)\right)\mathbf{V}\cdot\nabla\chi\beta_\theta(u)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \end{aligned}$$

verifies

$$F(u,\chi) \leq 0, \quad \forall\,\chi \in \mathcal{C}^1\left([0,T); H_0^1(\Omega)\right) \text{ with } \chi(\cdot,T) = 0 \text{ and } \chi \geq 0. \tag{5.6}$$

Furthermore, it is required to satisfy

$$\begin{aligned} &\forall\varepsilon > 0, \exists\, Q^\varepsilon \subset Q_T \text{ such that meas}\left(Q^\varepsilon\right) < \varepsilon, \text{ and} \\ &F(u,\chi) = 0, \quad \forall\,\chi \in \mathcal{C}^1\left([0,T); H_0^1(\Omega)\right), \text{supp}\,\chi \subset \left([0,T)\times\Omega\right)\setminus Q^\varepsilon \end{aligned} \tag{5.7}$$

Theorem 5.4. *Under the assumptions* (H1) − (H3) *and* (H4b)*, there exists at least one weak degenerate solution to system* (5.1)–(5.3) *in the sense of Definition 5.3.*

We give now the definition of weak solutions to system (5.1)–(5.3) when the assumption (H4c) is satisfied. Let $\theta, \lambda \geq 0$, we denote by $j_{\theta,\lambda}$ the continuous function defined by

$$j_{\theta,\lambda}(u) = \begin{cases} \beta_\theta(u), & \text{if } 0 \leq u \leq u_* \\ \beta_\theta(u_*)\,(1-u_*)^{1-\frac{r'}{2}-\lambda}(1-u)^{\frac{r'}{2}-1} + \lambda, & \text{if } u \geq u_*. \end{cases} \tag{5.8}$$

where $r' \geq \max(2, r_2)$. We denote by $J_{\theta,\lambda}$ the primitive of the function $j_{\theta,\lambda}$

$$J_{\theta,\lambda} = \int_0^u j_{\theta,\lambda}(y)\,\mathrm{d}y. \tag{5.9}$$

To simplify the notations, we denote by $j = j_{0,0}$ and $J = J_{0,0}$.
In addition, we consider the functions μ and G, defined by

$$\begin{cases} \mu(u) = \beta(u), & 0 \le u \le u_* \\ \mu'(u) = (f(u) - g(u))^{-1} g'(u)\, \mu(u), & u_* \le u < 1. \end{cases} \tag{5.10}$$

G is the primitive of μ, that is

$$G(u) = \int_0^u \mu(y)\, \mathrm{d}y. \tag{5.11}$$

Definition 5.5. For $\theta \ge 7r_1 + 6 - r$, $\lambda \ge 7r_2 + 6 - \frac{r'}{2}$, and under the assumptions (H1)–(H3) and (H4c). We say that u is a degenerate weak solution to system (5.1)–(5.3) if

$$\begin{aligned} &0 \le u(\mathbf{x}, t) \le 1 \text{ for a.e. } (\mathbf{x}, t) \in \Omega \times (0, T), \\ &J(u) \in L^2\big(0, T; H_0^1(\Omega)\big), \quad \mu'^{\frac{1}{2}}(u)\, a^{\frac{1}{2}}(u)\, \nabla u \in \big(L^2(Q_T)\big)^d, \end{aligned}$$

and such that, the function F defined by

$$\begin{aligned} F(u, \chi) = &- \int_{Q_T} J_{\theta,\lambda}(u)\, \partial_t \chi \, \mathrm{d}\mathbf{x}\, \mathrm{d}t - \int_\Omega J_{\theta,\lambda}(u_0(\mathbf{x}))\, \chi(\mathbf{x}, 0)\, \mathrm{d}\mathbf{x} \\ &+ \int_{Q_T} a(u)\, \nabla u \cdot \nabla (j_{\theta,\lambda}(u)\, \chi)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t + \int_{Q_T} a(u)\, \mathbf{V} \cdot \nabla u\, j_{\theta,\lambda}(u)\, \chi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &- \int_{Q_T} (f(u) - g(u))\, \mathbf{V} \cdot \nabla (j_{\theta,\lambda}(u)\, \chi)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t + \int_{Q_T} g'(u)\, \mathbf{V} \cdot \nabla u\, j_{\theta,\lambda}(u)\, \chi\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &- \int_{Q_T} (f(u) - g(u))\, \mathbf{V} \cdot \nabla \chi\, j_{\theta,\lambda}(u)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t, \end{aligned}$$

verifies

$$F(u, \chi) \le 0, \quad \forall \chi \in \mathcal{C}^1\big([0, T); H_0^1(\Omega)\big) \text{ with } \chi(\cdot, T) = 0 \text{ and } \chi \ge 0 \tag{5.12}$$

and furthermore,

$$\begin{aligned} &\forall \varepsilon > 0, \exists\, Q^\varepsilon \subset Q_T \text{ such that meas}(Q^\varepsilon) < \varepsilon, \text{ and} \\ &F(u, \chi) = 0, \quad \forall \chi \in \mathcal{C}^1\big([0, T); H_0^1(\Omega)\big), \ \operatorname{supp} \chi \subset ([0, T) \times \Omega) \setminus Q^\varepsilon \end{aligned} \tag{5.13}$$

Theorem 5.6. *Under the assumptions* (H1) − (H3) *and* (H4c)*, there exists at least one degenerate weak solution to system* (5.1)–(5.3) *in the sense of Definition 5.5.*

Unless stated otherwise, κ represents a "generic" nonnegative quantity which need not have the same value through the proofs. Furthermore, C_α represents a nonnegative constant depending only on the subscript α.

5.3 Existence for the nondegenerate case

In this section, we prove the existence of solutions to the nondegenerate problem. To avoid the degeneracy of the dissipation function a, we introduce the modified problem where the dissipation a is replaced by $a_\eta(u) = a(u) + \eta$ in equation (5.1), with $0 < \eta \ll 1$.
Therefore, we consider the nondegenerate system

$$\partial_t u_\eta - \operatorname{div}\,(a_\eta(u_\eta)\, \nabla u_\eta - f(u_\eta)\mathbf{V}) - g(u_\eta) \operatorname{div}\,(\mathbf{V}) + a_\eta(u_\eta)\, \nabla u_\eta \cdot \mathbf{V} = 0, \text{ in } Q_T, \tag{5.14}$$

$$u_\eta(\mathbf{x}, t) = 0, \ \mathbf{V} \cdot \mathbf{n} \le 0, \quad \text{in } \Sigma_T, \tag{5.15}$$

$$u_\eta(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \text{in } \Omega. \tag{5.16}$$

5.3.1 Weak nondegenerate solutions

For the existence of a solution to the nondegenerate system, we make the following weak assumption

(H5) $a \in \mathcal{C}^0([0,1])$, $a(u) > 0$ for $0 < u < 1$, $a(0) = 0$, and $a(1) = 0$ or $a(1) > 0$.

Theorem 5.7 (nondegenerate system)**.** *For any fixed $\eta > 0$ and under the assumptions* (H1)$-$(H3) *and* (H5)*, there exists at least one weak solution u_η to the system* (5.14)–(5.16) *satisfying*

$$\begin{aligned} &0 \le u_\eta(\mathbf{x},t) \le 1 \quad \text{for a.e. } (\mathbf{x},t) \in Q_T, \\ &u_\eta \in L^2\big(0,T;H_0^1(\Omega)\big) \cap L^\infty\big(0,T;L^2(\Omega)\big), \\ &\partial_t u_\eta \in L^2\big(0,T;H^{-1}(\Omega)\big), \\ &u_\eta \in \mathcal{C}^0\big([0,T];L^2(\Omega)\big), \end{aligned} \tag{5.17}$$

and such that

$$\begin{aligned} \int_0^T \langle \partial_t u_\eta, \varphi\rangle \,\mathrm{d}\tau &+ \int_{Q_T} a_\eta(u_\eta)\nabla u_\eta \cdot \nabla\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_T} f(u_\eta)\mathbf{V}\cdot\nabla\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \int_{Q_T} g'(u_\eta)\mathbf{V}\cdot\nabla u_\eta\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T} g(u_\eta)\mathbf{V}\cdot\nabla\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \int_{Q_T} a_\eta(u_\eta)\mathbf{V}\cdot\nabla u_\eta\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0, \quad \forall\varphi \in L^2\big(0,T;H_0^1(\Omega)\big), \end{aligned} \tag{5.18}$$

where the bracket $\langle\cdot,\cdot\rangle$ represents the duality product between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Proof. The solutions to system (5.14)–(5.16) depend on the parameter η. To simplify the notations and for simplicity, we omit the dependence of solutions on the parameter η and we use u instead of u_η in this section. We will apply the Schauder fixed-point theorem to prove the existence of weak solutions to system (5.14)–(5.16).
It is necessary to use the continuous extension for the functions depending on u. For instance, we take $f(u) = g(u) = 1$ for all $u \ge 1$ and $f(u) = g(u) = 0$ for all $u \le 0$. Furthermore, we extend the dissipation a outside $[0,1]$ by taking

$$a(u) = 0, \text{ for u} \le 0, \quad \text{and} \quad a(u) = a(1), \text{ for u} \ge 0.$$

Fixed-point method

Let us introduce the closed subset $\mathcal{K}$ of $L^2(Q_T)$ given by

$$\mathcal{K} = \left\{ u \in L^2(Q_T);\ \|u\|^2_{L^\infty(0,T;L^2(\Omega))} + \eta\|u\|^2_{L^2\left(0,T;H_0^1(\Omega)\right)} \le A,\ \|\partial_t u\|_{L^2(0,T;H^{-1}(\Omega))} \le B \right\},$$

The constants A and B will be fixed later. The set $\mathcal{K}$ is a compact convex of $L^2\big(0,T;L^2(\Omega)\big)$(The compactness is due to the Aubin–Simon theorem [73]).
Let $\mathcal{T}$ be a map from $L^2\big(0,T;L^2(\Omega)\big)$ to $L^2\big(0,T;L^2(\Omega)\big)$ defined by

$$\mathcal{T}(\overline{u}) = u,$$

where u is the unique solution to the following **linear** parabolic equation

$$\partial_t u - \operatorname{div}\left(a_\eta(\overline{u})\nabla u - f(\overline{u})\mathbf{V}\right) - g(\overline{u})\operatorname{div}(\mathbf{V}) + a_\eta(\overline{u})\nabla u\cdot\mathbf{V} = 0 \tag{5.19}$$

with the associate initial and boundary conditions. The existence of a unique solution to problem (5.19) is obtained using the Galerkin method [56, 28]. Indeed, there exists a unique solution u to problem (5.19) verifying : $u \in L^2\left(0,T;H_0^1(\Omega)\right) \cap \mathcal{C}^0\left(0,T;L^2(\Omega)\right)$, $\partial_t u \in L^2\left(0,T;H^{-1}(\Omega)\right)$ such that, we have the following weak formulation : $\forall \varphi \in L^2\big(0,T;H_0^1(\Omega)\big)$,

$$\begin{aligned}\int_0^T \langle \partial_t u, \varphi\rangle \,\mathrm{d}t + \int_{Q_T} a_\eta(\overline{u})\,\nabla u\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_T} f(\overline{u})\,\mathbf{V}\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ + \int_{Q_T} g(\overline{u})\,\mathbf{V}\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T} a_\eta(\overline{u})\,\nabla u\cdot\mathbf{V}\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0.\end{aligned} \tag{5.20}$$

Lemma 5.8. *$\mathcal{T}$ is an application from $\mathcal{K}$ to $\mathcal{K}$.*

Proof. Since $u \in L^2\left(0,T;H_0^1(\Omega)\right)$, one takes the solution u as a test function in the weak formulation (5.20), and gets, for all $t \in (0,T)$, that

$$E_1 + E_2 = E_3 + E_4, \tag{5.21}$$

where

$$\begin{aligned}E_1 &= \frac{1}{2}\int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega |u(\mathbf{x},\tau)|^2 \mathrm{d}\mathbf{x}\right)\mathrm{d}\tau = \frac{1}{2}\|u(t)\|^2_{L^2(\Omega)} - \frac{1}{2}\|u_0\|^2_{L^2(\Omega)},\\ E_2 &= \int_0^t\int_\Omega a_\eta(\overline{u})\,\nabla u\cdot\nabla u\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau,\\ E_3 &= \int_0^t\int_\Omega \left(f(\overline{u}) - g(\overline{u})\right)\nabla u\cdot\mathbf{V}\mathrm{d}\mathbf{x}\,\mathrm{d}\tau,\\ E_4 &= -\int_0^t\int_\Omega a_\eta(\overline{u})\,\nabla u\cdot\mathbf{V}u\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau.\end{aligned}$$

From the degeneracy of the dissipation function a as well as its continuous extension, one has

$$E_2 = \int_0^t\int_\Omega \left(a(\overline{u}) + \eta\right)\nabla u\cdot\nabla u\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau \geq \eta\int_0^t\int_\Omega \nabla u\cdot\nabla u\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau. \tag{5.22}$$

Now, relying on the continuous extension of the functions f and g, the Cauchy-Schwarz, and the weighted Young inequality, one gets

$$\begin{aligned}|E_3| &\leq \int_0^t\int_\Omega |f(\overline{u}) - g(\overline{u})|\,|\nabla u\cdot\mathbf{V}|\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau \leq C_{f,g}\int_0^t\int_\Omega |\nabla u\cdot\mathbf{V}|\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau\\ &\leq C_{f,g}\,\|\nabla u\|_{(L^2(Q_t))^d}\cdot\|\mathbf{V}\|_{(L^2(Q_t))^d} \leq \kappa\,\|\nabla u\|^2_{(L^2(Q_t))^d} + \frac{C_{f,g,Q_t}}{4\kappa}\,\|\mathbf{V}\|^2_{(L^\infty(Q_t))^d},\end{aligned} \tag{5.23}$$

where κ is a constant to be specified later.
In the same manner, we have the following estimate

$$\begin{aligned}|E_4| &\leq \int_0^t\int_\Omega |a_\eta(\overline{u})|\,|\nabla u\cdot\mathbf{V}u|\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau \leq C_{a,\eta}\int_0^t\int_\Omega |\nabla u\cdot\mathbf{V}u|\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau\\ &\leq C_{a,\eta}\,\|\nabla u\|_{(L^2(Q_t))^d}\cdot\|\mathbf{V}u\|_{(L^2(Q_t))^d} \leq \kappa\,\|\nabla u\|^2_{(L^2(Q_t))^d} + \frac{C_{f,g,\mathbf{V}}}{4\kappa}\,\|u\|^2_{(L^2(Q_t))^d}.\end{aligned} \tag{5.24}$$

Choosing the constant $\kappa = \frac{\eta}{4}$ and plugging estimates (5.22)–(5.24) into equation (5.21) one can conclude that

$$\|u(t)\|^2_{L^2(\Omega)} + \eta\int_0^t\int_\Omega \nabla u\cdot\nabla u\,\mathrm{d}\mathbf{x}\,\mathrm{d}\tau \leq \kappa_1 + \kappa_2\,\|u\|^2_{L^2(Q_t)}, \tag{5.25}$$

or this is the same as

$$\|u(t)\|^2_{L^2(\Omega)} + \eta \|\nabla u\|^2_{(L^2(Q_t))^d} \leq \kappa_1 + \kappa_2 \int_0^t \|u(\tau)\|^2_{L^2(\Omega)} \, \mathrm{d}t, \tag{5.26}$$

where $\kappa_1 = \|u_0\|^2_{L^2(\Omega)} + \frac{C_{f,g,Q_t}}{2\kappa} \|\mathbf{V}\|^2_{(L^\infty(Q_t))^d}$ and $\kappa_2 = \frac{C_{f,g,\mathbf{V}}}{4\kappa}$.
From estimate (5.26), and thanks to the lemma of Grönwall, one can deduce that there exists a constant $\kappa_3 = (\kappa_1 \exp(\kappa_2 t)) > 0$ such that

$$\|u\|^2_{L^2(Q_t)} \leq k_3, \qquad \forall t \in (0,T). \tag{5.27}$$

Plugging estimate (5.27) into estimate (5.25), one has

$$\|u(t)\|^2_{L^2(\Omega)} + \eta \|\nabla u\|^2_{(L^2(Q_t))^d} \leq A, \qquad \forall t \in (0,T),$$

where $A = \kappa_1 + \kappa_2 \kappa_3$.
Consequently, one deduces that

$$\|u\|^2_{L^\infty(0,T;L^2(\Omega))} + \eta \|u\|^2_{L^2\left(0,T;H^1_0(\Omega)\right)} \leq A.$$

It remains to show the estimate on $\partial_t u$. To do this, we take $\varphi \in L^2\left(0,T;H^1_0(\Omega)\right)$ as a test function into the weak formulation (5.20), one gets

$$\begin{aligned}
\left| \int_0^T \langle \partial_t u, \varphi \rangle \, \mathrm{d}t \right| &\leq \int_{Q_T} |f(\overline{u}) - g(\overline{u})| \, |\mathbf{V} \cdot \nabla \varphi| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t + \int_{Q_T} |a_\eta \overline{u}| \, |\nabla u \cdot (\nabla \varphi + \mathbf{V}\varphi)| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\
&\leq C_{f,g} \|\mathbf{V}\|_{(L^2(Q_T))^d} \|\nabla \varphi\|_{(L^2(Q_T))^d} + C_{a,\eta} \|\nabla u\|_{(L^2(Q_T))^d} \|\nabla \varphi\|_{(L^2(Q_T))^d} \\
&\quad + C_{a,\eta,\mathbf{V}} \|\nabla u\|_{(L^2(Q_T))^d} \|\varphi\|_{L^2(Q_T)}.
\end{aligned}$$

Note that the Poincaré inequality implies the existence of a constant $\kappa_4 > 0$ (depending only on the domain Ω) such that

$$\|\varphi\|_{L^2(Q_T)} \leq \kappa_4 \|\nabla \varphi\|_{(L^2(Q_T))^d}.$$

Therefore, one can deduce that

$$\left| \int_0^t \langle \partial_t u, \varphi \rangle \, \mathrm{d}t \right| \leq B \|\nabla \varphi\|_{(L^2(Q_T))^d}.$$

This ends the proof of the lemma. □

Lemma 5.9. *$\mathcal{T}$ is a continuous application.*

Proof. Let $(\overline{u}_n)_n$ be a sequence of $\mathcal{K}$ and $\overline{u} \in \mathcal{K}$ such that $\overline{u}_n \longrightarrow \overline{u}$ converges strongly in $L^2\left(0,T;L^2(\Omega)\right)$. In order to prove the lemma, it suffices to show that

$$\mathcal{T}(\overline{u}_n) = u_n \longrightarrow \mathcal{T}(\overline{u}) = u \text{ converges strongly in } L^2\left(0,T;L^2(\Omega)\right).$$

For all $\varphi \in L^2\left(0,T;H^1_0(\Omega)\right)$, the sequence $(u_n)_n$ satisfies

$$\begin{aligned}
&\int_0^T \langle \partial_t u_n, \varphi \rangle \, \mathrm{d}t + \int_{Q_T} a_\eta(\overline{u}_n) \nabla u_n \cdot \nabla \varphi \, \mathrm{d}\mathbf{x} \, \mathrm{d}t - \int_{Q_T} f(\overline{u}_n) \mathbf{V} \cdot \nabla \varphi \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\
&\qquad + \int_{Q_T} g(\overline{u}_n) \mathbf{V} \cdot \nabla \varphi \, \mathrm{d}\mathbf{x} \, \mathrm{d}t + \int_{Q_T} a_\eta(\overline{u}_n) \mathbf{V} \cdot \nabla u_n \varphi \, \mathrm{d}\mathbf{x} \, \mathrm{d}t = 0.
\end{aligned} \tag{5.28}$$

Let us denote v_n by $v_n = u_n - u$ and substrat equation (5.20) from equation (5.28), one has $\forall \varphi \in L^2\left(0,T;H_0^1(\Omega)\right)$

$$\begin{aligned}\int_0^T \langle \partial_t v_n, \varphi\rangle \,\mathrm{d}t &+ \int_{Q_T} a_\eta(\overline{u}_n)\nabla v_n\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &- \int_{Q_T}\left(f(\overline{u}_n)-f(\overline{u})\right)\mathbf{V}\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left(g(\overline{u}_n)-g(\overline{u})\right)\mathbf{V}\cdot\nabla\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &+ \int_{Q_T} a_\eta(\overline{u}_n)\mathbf{V}\cdot\nabla v_n\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\cdot\mathbf{V}\varphi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0.\end{aligned} \tag{5.29}$$

Now, we take $\varphi = v_n$ as a test function in equation (5.29), and a parameter $\delta > 0$ defined later, we have the following equation

$$\sum_{1\le i\le 7} H_i = 0,$$

where

$$\begin{aligned}H_1 &= \int_0^t \langle \partial_t v_n, v_n\rangle\,\mathrm{d}\tau = \frac{1}{2}\left\|v_n(t)\right\|^2_{L^2(\Omega)},\\ H_2 &= \int_{Q_t} a_\eta(\overline{u}_n)\nabla v_n\cdot\nabla v_n\,\mathrm{d}\tau\mathrm{d}\mathbf{x} \ge \eta\left\|\nabla v_n\right\|^2_{(L^2(Q_t))^d},\\ H_3 &= \left|\int_{Q_t}\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\cdot\nabla v_n\,\mathrm{d}\tau\mathrm{d}\mathbf{x}\right|\\ &\le \delta\left\|\nabla v_n\right\|^2_{(L^2(Q_t))^d} + \frac{1}{4\delta}\left\|\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\right\|^2_{(L^2(Q_t))^d},\\ H_4 &= \left|\int_{Q_t}\left(f(\overline{u}_n)-f(\overline{u})\right)\nabla v_n\cdot\mathbf{V}\,\mathrm{d}\tau\mathrm{d}\mathbf{x}\right|\\ &\le \delta\left\|\nabla v_n\right\|^2_{(L^2(Q_t))^d} + \frac{1}{4\delta}\left\|\left(f(\overline{u}_n)-f(\overline{u})\right)\mathbf{V}\right\|^2_{(L^2(Q_t))^d},\\ H_5 &= \left|\int_{Q_t}\left(g(\overline{u}_n)-g(u_n)\right)\nabla v_n\cdot\mathbf{V}\,\mathrm{d}\tau\mathrm{d}\mathbf{x}\right|\\ &\le \delta\left\|\nabla v_n\right\|^2_{(L^2(Q_t))^d} + \frac{1}{4\delta}\left\|\left(g(\overline{u}_n)-g(\overline{u})\right)\mathbf{V}\right\|^2_{(L^2(Q_t))^d},\\ H_6 &= \left|\int_{Q_t} a_\eta(\overline{u}_n)\nabla v_n\cdot\mathbf{V}v_n\,\mathrm{d}\tau\mathrm{d}\mathbf{x}\right| \le \delta\left\|\nabla v_n\right\|^2_{(L^2(Q_t))^d} + \frac{C_{a,\eta}}{4\delta}\left\|v_n\right\|^2_{L^2(Q_t)},\\ H_7 &= \left|\int_{Q_t}\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\cdot\mathbf{V}v_n\,\mathrm{d}\tau\mathrm{d}\mathbf{x}\right|\\ &\le \delta\left\|\nabla v_n\right\|^2_{(L^2(Q_t))^d} + C_{a,\eta,\mathbf{V}}\left\|\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\right\|^2_{(L^2(Q_t))^d}.\end{aligned}$$

Plugging these estimates into equation (5.29) and choosing $\delta = \frac{\eta}{12}$, one can deduce that

$$\left\|v_n(t)\right\|^2_{L^2(\Omega)} \le \kappa_5 + \frac{6C_{a,\eta}}{\eta}\left\|v_n\right\|^2_{L^2(Q_t)},$$

where

$$\begin{aligned}\kappa_5 = \frac{1}{2\delta}\Big(&\left\|\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\right\|^2_{(L^2(Q_t))^d} + \left\|\left(f(\overline{u}_n)-f(\overline{u})\right)\mathbf{V}\right\|^2_{(L^2(Q_t))^d}\\ &\left\|\left(g(\overline{u}_n)-g(\overline{u})\right)\mathbf{V}\right\|^2_{(L^2(Q_t))^d} + C_{a,\eta,\mathbf{V}}\left\|\left(a_\eta(\overline{u}_n)-a_\eta(\overline{u})\right)\nabla u\right\|^2_{(L^2(Q_t))^d}\Big).\end{aligned}$$

Now, thanks to the Grönwall lemma, one can deduce that

$$\|v_n(t)\|^2_{L^2(\Omega)} \leq \kappa_5 \exp\left(\frac{6C_{a,\eta}t}{\eta}\right).$$

Note that κ_5 tends to zero as $n \to \infty$. Indeed, one can get the result using Lebesgue's dominated convergence theorem, the uniqueness of a solution to problem (5.19), and using the continuity of each of the functions a, f, and g. Therefore, one gets

$$\|v_n(t)\|_{L^2(\Omega)} \xrightarrow[n\to+\infty]{} 0, \qquad \forall t \in (0,T).$$

In other words, one has

$$\|v_n\|_{L^\infty(0,T;L^2(\Omega))} \longrightarrow 0 \text{ quand } n \longrightarrow +\infty,$$

which implies that

$$u_n \longrightarrow u \text{ strongly in } L^2\left(0,T;L^2(\Omega)\right).$$

This ends the proof of this lemma. □

Using previous results, Green-Riemann's theorem, and Schauder's fixed-point theorem, one can deduce that there exists at least one solution to the nondegenerate problem (5.14)–(5.16) in the sense of theorem 5.7. It remains to show that the solution verifies the maximum principle.

5.3.2 Maximum principle on the saturation

In this section, we aim to prove that the solution to the nondegenerate problem (5.14)–(5.16) is stable in the sense of verifying the maximum principle. Specifically, we have the following lemma.

Lemma 5.10. *Let u be a solution to the nondegenerate system (5.14)–(5.16) under the assumptions (H1) − (H3) and (H5). Then, the solution u satisfies*

$$0 \leq u(\mathbf{x},t) \leq 1, \quad \text{for a.e. } (\mathbf{x},t) \in Q_T.$$

Proof. Let u^- be the function defined by $u^- = \max(-u,0) = \frac{|u|-u}{2} \geq 0$. Stampacchia's Theorem ensures that $u^- \in L^2\left(0,T;H^1_0(\Omega)\right)$ since $u \in L^2\left(0,T;H^1_0(\Omega)\right)$. Therefore, one can consider $-u^-$ as a test function into the weak formulation (5.20), and gets using the inequality $u = u^+ - u^-$ that

$$\begin{aligned}\int_0^t \langle \partial_t u^-, u^- \rangle \,\mathrm{d}t + \int_{Q_t} a_\eta(u)\nabla u^- \cdot \nabla u^- \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_t} f(u)\mathbf{V}\cdot\nabla u^- \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ + \int_{Q_t} g(u)\,\mathrm{div}\,(\mathbf{V})\,u^- \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_t} a_\eta(u)\nabla u^- \cdot \mathbf{V}u^- \,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0.\end{aligned} \tag{5.30}$$

For the first term of equation (5.30), one has

$$\int_0^t \langle \partial_t u^-, u^- \rangle \,\mathrm{d}t = \frac{1}{2}\int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\left\|u^-(\tau)\right\|^2_{L^2(\Omega)} \,\mathrm{d}\tau = \frac{1}{2}\left\|u^-(t)\right\|^2_{L^2(\Omega)},$$

since $u_0^- = 0$ due to the nonnegativity of the function $u_0(\mathbf{x})$ for a.e. $\mathbf{x} \in \Omega$.
Now, we use the definition of the function a_η and the degeneracy of the dissipation a to conclude that

$$\int_{Q_t} a_\eta(u)\nabla u^- \cdot \nabla u^- \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \geq \eta \int_{Q_t} \nabla u^- \cdot \nabla u^- \mathrm{d}\mathbf{x}\,\mathrm{d}t = \eta\left\|\nabla u^-\right\|^2_{(L^2(Q_t))^d}. \tag{5.31}$$

Furthermore, we rely on the continuous extension by zero of the functions $f(u)$ and $g(u)$ for $u \leq 0$, to deduce that the third and the fourth terms in equation (5.30) are equal to zero.
Let us now focus on the last term of equation (5.30). Indeed, by the Cauchy-Schwarz inequality as well as the weighted Young inequality, one has

$$\int_{Q_t} a_\eta(u) \nabla u^- \cdot \mathbf{V} u^- \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq \frac{\eta}{2} \left\|\nabla u^-\right\|^2_{(L^2(Q_t))^d} + \frac{C_{a,\eta,\mathbf{V}}}{2} \int_0^t \left\|u^-(\tau)\right\|^2_{L^2(\Omega)} \,\mathrm{d}\tau. \tag{5.32}$$

Substituting estimates (5.31)–(5.32) into equation (5.30), this yields

$$\left\|u^-(t)\right\|^2_{L^2(\Omega)} + \eta \left\|\nabla u^-\right\|^2_{(L^2(Q_t))^d} \leq C_{a,\eta,\mathbf{V}} \int_0^t \left\|u^-(\tau)\right\|^2_{L^2(\Omega)} \,\mathrm{d}\tau.$$

Denote by ϕ the function defined by $\phi(t) = \int_0^t \|u^-(\tau)\|^2_{L^2(\Omega)} \,\mathrm{d}\tau$, then one has

$$\frac{\mathrm{d}\phi}{\mathrm{d}t}(t) \leq C_{a,\eta,\mathbf{V}} \phi(t).$$

Applying, the Grönwall lemma, one can deduce that

$$\left\|u^-(t)\right\|^2_{L^2(\Omega)} = \phi(t) \leq \phi(0) \exp\left(C_{a,\eta,\mathbf{V}} t\right) = 0.$$

As a consequence, $u^-(\mathbf{x},t) = 0$, for a.e. $(\mathbf{x},t) \in Q_T$, i.e. $u(\mathbf{x},t) \geq 0$, for a.e. $(\mathbf{x},t) \in Q_T$.
It remains to show that $u(\mathbf{x},t) \leq 1$, for a.e. $(\mathbf{x},t) \in Q_T$. To do this, it suffices to prove that $(u-1)^+ = 0$. Thus, we multiply the saturation equation (5.14) by the regular function $(u-1)^+ \in L^2\left(0,T;H^1_0(\Omega)\right)$ and integrate the resulting equation over $\Omega \times (0,t)$, this yields

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \int_0^t \left\|(u-1)^+(\tau)\right\|^2_{L^2(\Omega)} \mathrm{d}\tau + \int_{Q_t} a_\eta(u) \nabla (u-1)^+ \cdot \nabla (u-1)^+ \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad - \int_{Q_t} f(u) \mathbf{V} \cdot \nabla (u-1)^+ \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_t} g(u) \,\mathrm{div}\,(\mathbf{V}) (u-1)^+ \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad + \int_{Q_t} a_\eta(u) \nabla u \cdot \mathbf{V} (u-1)^+ \,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0. \end{aligned} \tag{5.33}$$

Now, we proceed as before and get the estimates for each term of equation (5.33).
For the first term, one has

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \int_0^t \left\|(u-1)^+(\tau)\right\|^2_{L^2(\Omega)} \mathrm{d}\tau = \frac{1}{2} \left\|(u-1)^+(t)\right\|^2_{L^2(\Omega)} - \frac{1}{2} \left\|(u_0-1)^+\right\|^2_{L^2(\Omega)}, \tag{5.34}$$

and since $u_0 \leq 1$ then the second term on the right-hand side of (5.34) is equal to zero.
For the third and the fourth term of equation (5.33), by using the fact that $f(u) = g(u) = 1$ for all $u \geq 1$ and the fact that $\mathbf{V} \cdot \mathbf{n} \leq 0$ on $\partial\Omega$, one has

$$\begin{aligned} &- \int_{Q_t} f(u) \mathbf{V} \cdot \nabla (u-1)^+ \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_t} g(u) \,\mathrm{div}\,(\mathbf{V}) (u-1)^+ \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad = - \int_{\Sigma_T} (u-1)^+ \mathbf{V} \cdot \mathbf{n} \,\mathrm{d}\sigma(\mathbf{x})\,\mathrm{d}t \geq 0 \end{aligned}$$

Finally, for the last term of equation (5.34), we use again the extension by $a(1)$ of the dissipation function a for $u > 1$, the Cauchy-Schwarz inequality and the weighted Young inequality, and get

the following estimate

$$\int_{Q_t} a_\eta(u)\nabla u\cdot\mathbf{V}(u-1)^+\mathrm{d}\mathbf{x}\,\mathrm{d}t=\int_{Q_t} a_\eta(u)\nabla(u-1)\cdot\mathbf{V}(u-1)^+\mathrm{d}\mathbf{x}\,\mathrm{d}t$$
$$\leq\frac{\eta}{2}\left\|\nabla(u-1)^+\right\|^2_{(L^2(Q_t))^2}+\frac{C_{a,\eta,\mathbf{V}}}{2}\int_0^t\left\|(u-1)^+(\tau)\right\|^2_{L^2(\Omega)}\,\mathrm{d}\tau.$$

Plugging the previous estimates into equation (5.33), one has

$$\left\|(u-1)^+(t)\right\|^2_{L^2(\Omega)}+\eta\left\|\nabla(u-1)^+\right\|^2_{(L^2(Q_t))^d}\leq C_{a,\eta,\mathbf{V}}\int_0^t\left\|(u-1)^+(\tau)\right\|^2_{L^2(\Omega)}\,\mathrm{d}\tau.$$

One can conclude, using the Grönwall lemma, that $u(\mathbf{x},t)\leq 1$, for a.e. $(\mathbf{x},t)\in Q_T$. This ends the proof of lemma 5.10. □

The proof of theorem 5.7 is now completed. □

5.4 Proof of theorem 5.2.

In the previous section, we have shown that the nondegenerate system (5.14)–(5.16) admits at least one weak solution. Here, we are going to prove theorem (5.2), the proof is based on the establishment of estimates on the solutions independent of the parameter η, and next on the passage to the limit as η tends to zero.

From the definition (5.4) of k, we have

$$k(u)=k(u_*)\exp\left(\int_{u_*}^{u}\left(f(\tau)-g(\tau)\right)^{-1}g'(\tau)\right)\mathrm{d}\tau,\qquad\text{for all } u\geq u_*.$$

As a consequence of assumption (H4a), there exist two constants c_3 and c_4 depending on f, g, and u_* such that

$$c_3(1-u)^{-1}\leq k(u)\leq c_4(1-u)^{-1},\qquad\forall u_*\leq u<1. \tag{5.35}$$

Indeed, we have

$$k(u_*)\exp\left(c_1C_{g'}\int_{u_*}^{u_\eta}\frac{1}{1-\tau}\mathrm{d}\tau\right)\leq k(u_\eta\leq k(u_*)\exp\left(c_2\left\|g'\right\|_\infty\int_{u_*}^{u_\eta}\frac{1}{1-\tau}\mathrm{d}\tau\right).$$

That is

$$\frac{c_1C_{g'}k(u_*)(1-u_*)}{1-u_\eta}\leq k(u_\eta)\leq\frac{c_2\left\|g'\right\|_\infty k(u_*)(1-u_*)}{1-u_\eta}.$$

Denoting by $c_3=c_1C_{g'}k(u_*)(1-u_*)$ and $c_4=c_2||g'||_\infty k(u_*)(1-u_*)$, then one obtains the confinement (5.35).

Now, using the confinement of the function k and denoting by $c_5=c_1c_3C_{g'}$ and $c_6=c_2c_4\left\|g'\right\|_\infty$, one can easily obtain that

$$\frac{c_5}{(1-u_\eta)^{-2}}\leq k'(u_\eta)\leq\frac{c_6}{(1-u_\eta)^{-2}}.$$

Lemma 5.11. *Let $L(u_0)$ belongs to $L^1(\Omega)$. Under the assumptions (H1) − (H3) and (H4a), the weak solution to the nondegenerate system (5.14)–(5.16) verifies*

The sequence $(\nabla u_\eta)_\eta$ is uniformly bounded in $\left(L^2(Q_T)\right)^d$. (5.36)

The sequence $\left(\sqrt{\eta k'(u_\eta)}\nabla u_\eta\right)_\eta$ is uniformly bounded in $\left(L^2(Q_T)\right)^d$. (5.37)

The sequence $(\partial_t u_\eta)_\eta$ is uniformly bounded in $L^2\left(0,T;H^{-1}(\Omega)\right)$. (5.38)

The sequence $(u_\eta)_\eta$ relatively compact in $L^2(Q_T)$. (5.39)

Proof. To prove (5.36) and (5.37), we multiply the saturation equation (5.14) by k and integrate it over Ω, one gets

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega} L(u_\eta)\,\mathrm{d}\mathbf{x} &+ \int_{\Omega} a(u_\eta)\,k'(u_\eta)\,\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x} + \eta\int_{\Omega} k'(u_\eta)\,\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\\ &= \int_{\Omega}\left(f(u_\eta)-g(u_\eta)\right)\mathbf{V}\cdot\nabla k(u_\eta)\,\mathrm{d}\mathbf{x} - \int_{\Omega} g'(u_\eta)\,k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\\ &- \int_{\Omega} a(u_\eta)\,k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x} - \eta\int_{\Omega} k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}.\end{aligned}\tag{5.40}$$

We denote $\Omega_1=\Omega\cap\{u_\eta<u_*\}$ and $\Omega_2=\Omega\cap\{u_\eta\geq u_*\}$; then the whole integral on Ω can be split into two sub integrals on Ω_1 and Ω_2 respectively. Within the region Ω_1, we have $a(u_\eta)\geq a_0$ and $k'(u_\eta)=1$, this yields

$$\int_{\Omega_1} a(u_\eta)\,k'(u_\eta)\,\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\geq a_0\int_{\Omega_1}\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}=a_0\left\|\nabla u_\eta\right\|^2_{(L^2(\Omega))^d}.$$

Furthermore, we have the following estimates

$$\begin{aligned}\left|\int_{\Omega_1}\left(f(u_\eta)-g(u_\eta)\right)k'(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\right| &\leq \int_{\Omega_1}\left|\left(f(u_\eta)-g(u_\eta)\right)\mathbf{V}\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\\ &\leq \frac{a_0}{6}\left\|\nabla u_\eta\right\|^2_{(L^2(\Omega_1))^d}+\frac{3C}{2a_0}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega_1))^d},\\ \left|\int_{\Omega_1}\left(g'(u_\eta)+a(u_\eta)\right)k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\mathrm{d}\mathbf{x}\right| &\leq \int_{\Omega_1}\left|\left(g'(u_\eta)+a(u_\eta)\right)\mathbf{V}\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\\ &\leq \frac{a_0}{3}\left\|\nabla u_\eta\right\|^2_{(L^2(\Omega_1))^d}+\frac{3\left(\|g'\|_\infty+\|a\|_\infty\right)^2}{4a_0}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega_1))^d},\\ \left|\eta\int_{\Omega_1} k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\right| &\leq \int_{\Omega_1}\left|\eta\,u_\eta\mathbf{V}\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\leq\int_{\Omega_1}\left|\eta\mathbf{V}\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\\ &\leq \frac{1}{2}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega_1))^d}+\frac{1}{2}\int_{\Omega_1}\eta\,k'(u_\eta)\,\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}.\end{aligned}$$

In region Ω_2, we have

$$\begin{aligned}&\int_{\Omega_2}\left(\left(f(u_\eta)-g(u_\eta)\right)k'(u_\eta)-g'(u_\eta)\,k(u_\eta)\right)\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}=0,\\ &\int_{\Omega_2} a(u_\eta)\,k'(u_\eta)\,\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\geq c_5\int_{\Omega_2} a(u_\eta)\left(1-u_\eta\right)^{-2}\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}.\end{aligned}$$

The following estimates hold

$$\begin{aligned}\left|\int_{\Omega_2}\eta\,k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\right| &\leq c_4\int_{\Omega_2}\left|\sqrt{\eta}\left(1-u_\eta\right)^{-1}\mathbf{V}\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\\ &\leq \frac{c_4^2}{2c_5}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega))^d}+\frac{1}{2}\int_{\Omega_2}\eta\,k'(u_\eta)\,\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x},\\ \left|\int_{\Omega_2} a(u_\eta)\,k(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\right| &\leq c_4\int_{\Omega_2}\left|\sqrt{a(u_\eta)}\left(1-u_\eta\right)^{-1}\mathbf{V}\cdot\sqrt{a(u_\eta)}\nabla u_\eta\right|\mathrm{d}\mathbf{x}\\ &\leq c_4\left(\int_{\Omega_2} a(u_\eta)\,\mathbf{V}^2\mathrm{d}\mathbf{x}\right)^{\frac{1}{2}}\left(\int_{\Omega_2} a(u_\eta)\left(1-u_\eta\right)^{-2}\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}\right)^{\frac{1}{2}}\\ &\leq \frac{c_4^2\left\|a\right\|^2_\infty}{2c_5}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega))^d}+\frac{c_5}{2}\int_{\Omega_2} a(u_\eta)\left(1-u_\eta\right)^{-2}\nabla u_\eta\cdot\nabla u_\eta\,\mathrm{d}\mathbf{x}.\end{aligned}$$

From the degeneracy assumption (H4a), we have $a\left(u_\eta\right) \geq m_1(1-u_\eta)^{r_2}$ for an exponent r_2 such that $r_2 \leq 2$, and therefore

$$\int_{\Omega_2} a\left(u_\eta\right)\left(1-u_\eta\right)^{-2} \nabla u_\eta \cdot \nabla u_\eta \, \mathrm{d}\mathbf{x} \geq m_1(1-u_*)^{r_2-2} \int_{\Omega_2} \nabla u_\eta \cdot \nabla u_\eta \, \mathrm{d}\mathbf{x}.$$

Plugging the previous estimates into equation (5.40), this yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega L\left(u_\eta\right)\mathrm{d}\mathbf{x} + \frac{a_0}{2}\int_{\Omega_1} \nabla u_\eta \cdot \nabla u_\eta \mathrm{d}\mathbf{x} + \frac{c_5 m_1\left(1-u_*\right)^{r_2-2}}{2}\int_{\Omega_2} \nabla u_\eta \cdot \nabla u_\eta \, \mathrm{d}\mathbf{x} \\ + \frac{1}{2}\int_{\Omega_2} \eta\, k'\left(u_\eta\right) \nabla u_\eta \cdot \nabla u_\eta \, \mathrm{d}\mathbf{x} \leq C_{a,g',a_0},\end{aligned}$$

where, C_{a,g',a_0} is a nonnegative constant equals to

$$C_{a,g',a_0} = \frac{1}{2}\left(\frac{3C}{a_0} + \frac{3\left(\|g'\|_\infty + \|a\|_\infty\right)}{2a_0} + \frac{c_4^2\|a\|_\infty^2}{c_5} + 1 + \frac{c_4^2}{c_5}\right)\|\mathbf{V}\|^2_{(L^2(\Omega))^d}.$$

As a consequence, one has

$$2\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega L\left(u_\eta\right)\mathrm{d}\mathbf{x} + C_{a_0,u_*,m_1}\|\nabla u_\eta\|^2_{L^2(\Omega)} + \left\|\sqrt{\eta k'\left(u_\eta\right)}\nabla u_\eta\right\|^2_{L^2(\Omega)} \leq 2C_{a,g',a_0}, \tag{5.41}$$

where, $C_{a_0,u_*,m_1} = \min\left(a_0, c_5 m_1\left(1-u_*\right)^{r_2-2}\right)$.
Non, we integrate inequality (5.41) with respect to the time over $(0,t)\,,\ t \in (0,T)$, one gets

$$\begin{aligned}2\int_\Omega L\left(u_\eta\right)\mathrm{d}\mathbf{x} + C_{a_0,u_*}\|\nabla u_\eta\|^2_{L^2(Q_t)} + \left\|\sqrt{\eta k'\left(u_\eta\right)}\nabla u_\eta\right\|^2_{L^2(Q_t)} \\ \leq 2\left(TC_{a,g',a_0} + \|L\left(u_0\right)\|_{L^1(\Omega)}\right).\end{aligned}$$

One can conclude, using these estimate, that the sequences $\left(\nabla u_\eta\right)_\eta$ and $\left(\sqrt{\eta k'\left(u_\eta\right)}\nabla u_\eta\right)_\eta$ are uniformly bounded in $\left(L^2\left(0,T;L^2\left(\Omega\right)\right)\right)^d$.
Let us prove the third part (5.38) of this lemma, for that we take $\varphi \in L^2\left(0,T;H_0^1\left(\Omega\right)\right)$ as a test function into the weak formulation (5.18), one gets

$$\begin{aligned}\left|\int_0^T \left\langle \partial_t u_\eta, \varphi\right\rangle \mathrm{d}t\right| \leq \int_{Q_T}\left|a\left(u_\eta\right)\nabla u_\eta \cdot \nabla\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left|a\left(u_\eta\right)\mathbf{V}\cdot\nabla u_\eta\,\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ + \int_{Q_T}\left|\left(f\left(u_\eta\right) - g\left(u_\eta\right)\right)\mathbf{V}\cdot\nabla\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left|g'\left(u_\eta\right)\mathbf{V}\cdot\nabla u_\eta\,\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ + \int_{Q_T}\left|\eta\,\nabla u_\eta\cdot\nabla\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left|\eta\,\mathbf{V}\cdot\nabla u_\eta\,\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t.\end{aligned}$$

Using the Cauchy-Schwarz inequality as well as the previous estimates, one can deduce that

$$\begin{aligned}\int_{Q_T}\left|a\left(u_\eta\right)\nabla u_\eta\cdot\nabla\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left|\left(f\left(u_\eta\right) - g\left(u_\eta\right)\right)\mathbf{V}\cdot\nabla\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ + \int_{Q_T}\left|\eta\nabla u_\eta\cdot\nabla\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C_{a_0,a,u_*,g',L,\mathbf{V}}\|\nabla\varphi\|_{(L^2(Q_T))^d}.\end{aligned}$$

also, one gets

$$\int_{Q_T} \left|g'\left(u_\eta\right)\mathbf{V}\cdot\nabla u_\eta\,\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left|a\left(u_\eta\right)\mathbf{V}\cdot\nabla u_\eta\,\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T}\left|\eta\,\mathbf{V}\cdot\nabla u_\eta\,\varphi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq C_{a_0,a,u_*,g',L,\mathbf{V}}\|\varphi\|_{L^2(Q_T)}.$$

Using the Poincaré inequality, one deduces that there exists a constant $C_{a_0,a,u_*,g',L,\mathbf{V},\Omega,T} > 0$ such that

$$\left|\langle\partial_t u_\eta,\varphi\rangle\right| \leq C_{a_0,a,u_*,g',L,\mathbf{V},\Omega,T}\left\|\varphi\right\|_{L^2\left(0,T;H_0^1(\Omega)\right)}, \qquad \forall\varphi\in L^2\left(0,T;H_0^1\left(\Omega\right)\right).$$

This proves that the sequence $\left(\partial_t u_\eta\right)_\eta$ is uniformly bounded in $L^2\left(0,T;H^{-1}\left(\Omega\right)\right)$.
It remains to show the last part (5.39) of the lemma. Indeed, using statements (5.36) and (5.38), we remark that the sequence $\left(u_\eta\right)_\eta$ belongs to the Sobolev space

$$\mathcal{W} = \left\{u_\eta; u_\eta\in L^2\left(0,T;H_0^1\left(\Omega\right)\right) \text{ and } \partial_t u_\eta\in L^2\left(0,T;H^{-1}\left(\Omega\right)\right)\right\}.$$

Since $H_0^1\left(\Omega\right)$ is compactly embedded in $L^2\left(\Omega\right)$ and $L^2\left(\Omega\right)$ is continuously embedded in $H^{-1}\left(\Omega\right)$, then thanks to the Aubin–Lions lemma, $\mathcal{W}$ is compactly embedded in $L^2\left(0,T;L^2\left(\Omega\right)\right)$. Consequently, the sequence $\left(u_\eta\right)_\eta$ is relatively compact in $L^2\left(0,T;L^2\left(\Omega\right)\right)$. This ends the proof of lemma 5.11. □

The next step of the proof of theorem 5.2 is to pass to the limit as η tends to zero in the weak formulation (5.18).

Lemma 5.12. *We have the following convergences as η goes to zero*

$$u_\eta \longrightarrow u \text{ strongly in } L^2(Q_T), \tag{5.42}$$

$$\nabla u_\eta \longrightarrow \nabla u \text{ weakly in } \left(L^2\left(Q_T\right)\right)^d, \tag{5.43}$$

Furthermore,

$$u_\eta \longrightarrow u \text{ almost everywhere in } Q_T, \tag{5.44}$$

$$0\leq u\left(\mathbf{x},t\right)\leq 1 \text{ almost everywhere in } Q_T, \tag{5.45}$$

$$\partial_t u_\eta \longrightarrow \partial_t u \text{ weakly in } L^2\left(0,T;H^{-1}\left(\Omega\right)\right). \tag{5.46}$$

Proof. To prove the strong convergence (5.42), we deduce from lemma 5.11 and thanks to the Aubin–Lions lemma that there exists a subsequence also denoted $\left(u_\eta\right)_\eta$ and such that

$$u_\eta \longrightarrow u \text{ strongly in } L^2\left(Q_T\right).$$

On the other hand, and thanks to the uniform bound of the sequence $\left(\nabla u_\eta\right)_\eta$ in $\left(L^2\left(Q_T\right)\right)^d$, one can extract a subsequence such that

$$\nabla u_\eta \longrightarrow \nabla u \text{ weakly in } \left(L^2\left(Q_T\right)\right)^d,$$

to which the identification of the limit is made using the classical properties on the distribution theory.
The convergence (5.44) is a consequence of the strong convergence (5.42), while the convergence (5.45) is obtained using the fact that $\left\{u\in L^2\left(Q_T\right),\ 0\leq u\left(\mathbf{x},t\right)\leq 1 \text{ for a.e. } \left(\mathbf{x},t\right)\in Q_T\right\}$ is a closed subset of $L^2\left(Q_T\right)$.
Finally, the weak convergence (5.46) is a consequence of (5.38) and the strong convergence (5.42). □

Using the convergences given in lemma 5.12 as well as the Lebesgue theorem, one has for all $\varphi \in L^2\left(0,T;H_0^1(\Omega)\right)$

$$\int_{Q_T} \partial_t u_\eta \,\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \longrightarrow \int_{Q_t} \partial_t u \,\varphi \mathrm{d}\mathbf{x}\,\mathrm{d}t, \text{ as } \eta \to 0. \tag{5.47}$$

$$\int_{Q_T} a\left(u_\eta\right) \nabla u_\eta \cdot \nabla u_\eta \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \longrightarrow \int_{Q_T} a\left(u\right) \nabla u \cdot \nabla u \mathrm{d}\mathbf{x}\,\mathrm{d}t,, \text{ as } \eta \to 0. \tag{5.48}$$

$$\int_{Q_t} \left(f\left(u_\eta\right) - g\left(u_\eta\right)\right) \mathbf{V} \cdot \nabla \varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \longrightarrow \int_{Q_T} \left(f\left(u\right) - g\left(u\right)\right) \mathbf{V} \cdot \nabla \varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \text{ as } \eta \to 0. \tag{5.49}$$

Furthermore, one gets

$$\begin{aligned} &\int_{Q_T} \left(g'\left(u_\eta\right) + a\left(u_\eta\right)\right) \mathbf{V} \cdot \nabla u_\eta \,\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad \longrightarrow \int_{Q_T} \left(g'\left(u\right) + a\left(u\right)\right) \mathbf{V} \cdot \nabla u \,\varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \text{ as } \eta \to 0. \end{aligned} \tag{5.50}$$

It remains to show that

$$\int_{Q_T} \eta \nabla u_\eta \cdot \nabla \varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{Q_T} \eta \,\mathbf{V}\varphi \cdot \nabla u_\eta \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \longrightarrow 0, \text{ as } \eta \to 0.$$

Indeed, using the Cauchy-Schwarz inequality and the uniform bound of $\left(\nabla u_\eta\right)_\eta$, one has

$$\int_{Q_T} \eta \nabla u_\eta \cdot \nabla \varphi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le \eta \left\|\nabla u_\eta\right\|_{\left(L^2(Q_T)\right)^2} \left\|\nabla \varphi\right\|_{\left(L^2(Q_T)\right)^2} \longrightarrow 0, \text{ as } \eta \to 0. \tag{5.51}$$

$$\int_{Q_T} \eta \,\mathbf{V}\varphi \cdot \nabla u_\eta \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \le \eta\, C_{\mathbf{V}} \left\|\nabla u_\eta\right\|_{\left(L^2(Q_T)\right)^2} \left\|\nabla u_\eta\right\|_{L^2(Q_T)} \longrightarrow 0, \text{ as } \eta \to 0. \tag{5.52}$$

Using the convergence results (5.47)–(5.52), we can then pass to the limit in the weak formulation (5.18) and obtain the classical weak formulation on the limit solution u. This ends the proof of theorem 5.2.

5.5 Proof of theorem 5.4

In this section and under the degeneracy assumption (H4b) on the dissipation function a, we carry out some estimates on the solutions to the nondegenerate system (5.14)–(5.16) independent of η. We recall that the function β is defined as $\beta\left(u\right) = u^{r-1}$ and that h is its primitive defined by $h\left(u\right) = \int_0^u \beta\left(\mathbf{y}\right)\,\mathrm{d}\mathbf{y}$.

Lemma 5.13. *Under the assumptions* (H1) − (H3) *and* (H4b)*, the solutions to the nondegenerate system* (5.14)–(5.16) *satisfy*

$$0 \le u_\eta\left(\mathbf{x},t\right) \le 1, \textit{ for a.e. } \left(\mathbf{x},t\right) \in Q_T. \tag{5.53}$$

$$\left(\sqrt{u_\eta^{r-2} a\left(u_\eta\right)} \nabla u_\eta\right)_\eta \textit{ and } \left(u_\eta^{r-1} \nabla u_\eta\right)_\eta \textit{ are uniformly bounded in } \left(L^2\left(Q_T\right)\right)^d. \tag{5.54}$$

$$\left(\sqrt{\eta u_\eta^{r-2}} \nabla u_\eta\right)_\eta \textit{ and } \left(a\left(u_\eta\right) \nabla u_\eta\right)_\eta \textit{ are uniformly bounded in } \left(L^2\left(Q_T\right)\right)^d. \tag{5.55}$$

$$\left(h\left(u_\eta\right)\right)_\eta \textit{ is uniformly bounded in } L^\infty\left(0,T;L^2\left(\Omega\right)\right). \tag{5.56}$$

$$\left(\partial_t h\left(u_\eta\right)\right)_\eta \textit{ is uniformly bounded in } L^1\left(0,T;\left(W^{1,q}\left(\Omega\right)\right)'\right) \textit{ for } q > d. \tag{5.57}$$

$$\left(h\left(u_\eta\right)\right)_\eta \textit{ and } \left(u_\eta\right)_\eta \textit{ are relatively compact in } L^2\left(0,T;L^2\left(\Omega\right)\right). \tag{5.58}$$

Proof. The first part (5.53) is obtained in Lemma 5.10. Now, in order to obtain properties (5.54)–(5.56), we multiply the saturation equation (5.14) by $\beta(u_\eta)$, and get

$$E_1 + E_2 + E_3 + E_4 + E_5 + E_6 = 0. \tag{5.59}$$

where

$$\begin{aligned}
E_1 &= \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega h(u_\eta)\,\mathrm{d}\mathbf{x},\\
E_2 &= (r-1)\int_\Omega a(u_\eta)\,u_\eta^{r-2}\nabla u_\eta\cdot\nabla u_\eta \mathrm{d}\mathbf{x} = (r-1)\left\|\sqrt{u_\eta^{r-2}a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d},\\
E_3 &= \eta(r-1)\int_\Omega u_\eta^{r-2}|\nabla u_\eta|^2\mathrm{d}\mathbf{x} = (r-1)\left\|\sqrt{\eta u_\eta^{r-2}}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d},\\
E_4 &= (r-1)\int_\Omega (g(u_\eta)-f(u_\eta))\,u_\eta^{r-2}\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x} + \int_\Omega g'(u_\eta)\,u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x},\\
E_5 &= \int_\Omega a(u_\eta)\,u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}, \qquad E_6 = \eta\int_\Omega u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}.
\end{aligned}$$

Using assumptions (H2) and (H4b), the Cauchy-Schwarz and the weighted Young inequalities, we obtain the following estimate

$$\begin{aligned}
|E_4| \le{}& (r-1)\,C_{g'}\left(\int_\Omega \left|u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\right| + \int_\Omega \left|u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\right|\mathrm{d}\mathbf{x}\right)\\
\le{}& C_{m_1,r}\,\|\mathbf{V}\|^2_{(L^2(\Omega))^d} + \frac{m_1(r-1)}{4}\left\|u_\eta^{r-1}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d}.
\end{aligned}$$

In the same manner, we get the estimates for the last two terms

$$\begin{aligned}
|E_5| &\le \int_\Omega \left|a(u_\eta)\,u_\eta^{r-2}\nabla u_\eta\cdot\mathbf{V}\right|\mathrm{d}\mathbf{x} \le C_{a,r}\,\|\mathbf{V}\|^2_{(L^2(\Omega))^d} + \frac{(r-1)}{4}\left\|\sqrt{u_\eta^{r-2}a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d},\\
|E_6| &\le \int_\Omega \left|\eta\,u_\eta^{r-2}\nabla u_\eta\cdot\mathbf{V}\right|\mathrm{d}\mathbf{x} \le \frac{1}{r-1}\,\|\mathbf{V}\|^2_{(L^2(\Omega))^d} + \frac{(r-1)}{4}\left\|\eta^{\frac{1}{2}}u_\eta^{\frac{r-2}{2}}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d}.
\end{aligned}$$

Plugging the previous estimates into equation (5.59), this yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega h(u_\eta)\,\mathrm{d}\mathbf{x} + (r-1)\int_\Omega a(u_\eta)\,u_\eta^{r-2}\nabla u_\eta\cdot\nabla u_\eta\mathrm{d}\mathbf{x}&\\
+ (r-1)\int_\Omega \eta u_\eta^{r-2}\nabla u_\eta\cdot\nabla u_\eta\mathrm{d}\mathbf{x} &\le C_{m_1,r,a},
\end{aligned} \tag{5.60}$$

where $C_{m_1,r,a}$ is a nonnegative constant independent of η.
Thanks to statement (5.53), then assumption (H4b) ensures that $m_1 u_\eta^r \le m_1 u_\eta^{r_1} \le a(u_\eta) \le M_1 u_\eta^{r_1} \le M_1 u_\eta^{r-2}$, and therefore we have the following inequalities

$$\begin{aligned}
\left\|u_\eta^{r-1}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d} &\le \frac{1}{m_1}\left\|\sqrt{u_\eta^{r-2}a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d}, \text{ and}\\
\left\|a(u_\eta)\nabla u_\eta\right\|^2_{(L^2(\Omega))^d} &\le M_1\left\|\sqrt{u_\eta^{r-2}a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d}.
\end{aligned}$$

Integrating inequality (5.60) with respect to the time $t\in(0,T)$, then estimates (5.54)–(5.56) hold. To show that $(\partial_t h(u_\eta))_\eta$ is uniformly bounded in $L^1\left(0,T;\left(W^{-1}(\Omega)\right)'\right)$, we take a test function

$\chi \in L^2\left(0,T;H_0^1(\Omega)\right)\cap L^\infty(Q_T)$, multiply the saturation equation by $\beta(u_\eta)\chi$, and integrate the resulting equation over Q_T, one has

$$\begin{aligned}
&\int_0^T \langle \partial_t h(u_\eta), \chi\rangle \,\mathrm{d}t = -\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\nabla(\beta(u_\eta)\chi)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&-\eta\int_{Q_T}\nabla u_\eta\cdot\nabla(\beta(u_\eta)\chi)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t+\int_{Q_T}(f(u_\eta)-g(u_\eta))\mathbf{V}\cdot\nabla(\beta(u_\eta)\chi)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&-\int_{Q_T} g'(u_\eta)\nabla u_\eta\cdot\mathbf{V}\beta(u_\eta)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t-\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\mathbf{V}\beta(u_\eta)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&-\eta\int_{Q_T}\nabla u_\eta\cdot\mathbf{V}\beta(u_\eta)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.
\end{aligned}\tag{5.61}$$

Here, we give estimates on each integral on the right-hand side of equation (5.61) that we denote them $I_i, 1\le i\le 6$. To obtain the estimates, we use the Cauchy-Schwarz inequality. For the first term, we have

$$\begin{aligned}
|I_1| &\le \int_{Q_T}\left|a(u_\eta)\nabla u_\eta\cdot\left((r-1)u_\eta^{r-2}\nabla u_\eta\chi+u_\eta^{r-1}\nabla\chi\right)\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\le C_{r,a}\left(\int_{Q_T}\left|a(u_\eta)u_\eta^{r-2}\nabla u_\eta\cdot\nabla u_\eta\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t+\int_{Q_T}\left|u_\eta^{r-1}\nabla u_\eta\cdot\nabla\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\right)\\
&\le C_{r,a}\left\|\sqrt{a(u_\eta)u_\eta^{r-1}}\nabla u_\eta\right\|^2_{(L^2(Q_T))^d}\|\chi\|_{L^\infty(Q_T)}\\
&\qquad+\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\nabla\chi\|_{(L^2(Q_T))^d}.
\end{aligned}$$

In the same manner, we have the estimate on the second term

$$\begin{aligned}
|I_2| &\le \int_{Q_T}\left|\eta\nabla u_\eta\cdot\left((r-1)u_\eta^{r-2}\nabla u_\eta\chi+u_\eta^{r-1}\nabla\chi\right)\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\le C_{r,a}\left\|\sqrt{\eta u_\eta^{r-2}}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\chi\|_{L^\infty(Q_T)}\\
&\qquad+\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\nabla\chi\|_{(L^2(Q_T))^d}.
\end{aligned}$$

The third term is estimated, with the help of assumption (H2) and the Poincaré inequality, as follows

$$\begin{aligned}
|I_3| &\le \int_{Q_T}\left|(f(u_\eta)-g(u_\eta))\mathbf{V}\cdot\left((r-1)u_\eta^{r-2}\nabla u_\eta\chi+u_\eta^{r-1}\nabla\chi\right)\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\le C_{r,\Omega}\|\mathbf{V}\|^2_{(L^\infty(Q_T))^d}\left(\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}+1\right)\|\nabla\chi\|_{(L^2(Q_T))^d}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
|I_4| &\le \int_{Q_T}\left|g'(u_\eta)u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\le C_{g',\Omega}\left(\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\mathbf{V}\|^2_{(L^\infty(Q_T))^d}\right)\|\nabla\chi\|_{(L^2(Q_T))^d}.
\end{aligned}$$

Finally, the last two terms are estimated as

$$\begin{aligned}
|I_5+I_6| &\le \int_{Q_T}\left|a(u_\eta)u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t+\int_{Q_T}\left|\eta u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\le C_{a,\Omega}\left(\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\mathbf{V}\|^2_{(L^\infty(Q_T))^d}\right)\|\nabla\chi\|_{(L^2(Q_T))^d}.
\end{aligned}$$

Plugging the previous estimates into equation (5.61), one gets

$$|\langle \partial_t h(u_\eta), \chi \rangle| \leq \kappa \left(\|\chi\|_{L^\infty(Q_T)} + \|\chi\|_{L^2(0,T;H_0^1(\Omega))} \right).$$

One can conclude the proof of statement (5.57), using the boundedness of the sequences given in (5.53)–(5.55) and the embedding of the Sobolev space $W^{1,q}(\Omega) \subset H_0^1(\Omega) \cap L^\infty(\Omega)$ for $q > d$, and consequently, one has

$$L^\infty\left(0,T;W^{1,q}(\Omega)\right) \subset L^2\left(0,T;H^1(\Omega)\right) \cap L^\infty\left(0,T;L^2(\Omega)\right), \quad \forall q > d.$$

The last part (5.58) of the lemma is a direct consequence of the Aubin–Simon theorem. Indeed, one can prove that the sequence $(h(u_\eta))_\eta$ is uniformly bounded in $L^2\left(0,T;H_0^1(\Omega)\right)$ since $\nabla h(u_\eta) = u_\eta^{r-1}\nabla u_\eta$. Then the sequence $(h(u_\eta))_\eta$ is lying into the Sobolev space

$$\mathcal{W} = \left\{ u_\eta; u_\eta \in L^2\left(0,T;H_0^1(\Omega)\right) \text{ and } \partial_t u_\eta \in L^1\left(0,T;\left(W^{1,q}(\Omega)\right)'\right) \right\}.$$

Thanks to the Aubin–Lions lemma, $\mathcal{W}$ is compactly embedded in $L^2\left(0,T;L^2(\Omega)\right)$. Consequently, the sequence $(h(u_\eta))_\eta$ is relatively compact in $L^2\left(0,T;L^2(\Omega)\right)$.

Since the differentiable function h is nondecreasing, then h^{-1} exists and it is continuous, then the sequence $(u_\eta)_\eta$ is relatively compact in $L^2\left(0,T;L^2(\Omega)\right)$. The proof of lemma 5.13 is now complete. □

Now, we are interested in an almost-everywhere convergence of the gradient of the saturation weighted by a degenerate function of the saturation. Specifically, we have the following lemma.

Lemma 5.14. *Let $q = 3r_1 + 2$, where r_1 is defined in assumption* (H4b). *Then, the sequence $(u_\eta^q a(u_\eta)\nabla u_\eta)_\eta$ is a Cauchy sequence in measure.*

Proof. Let us denote by A, B, and b the functions defined by

$$A(u) = \int_0^u a(\tau)\,\mathrm{d}\tau, \quad B(u) = A^2(u), \quad b(u) = B'(u) = 2A(u)a(u). \tag{5.62}$$

For every $\mu > 0$ a nonnegative parameter, we consider the truncation function T_μ and its derivative Θ_μ defined by

$$T_\mu(u) = \min(\mu, \max(-\mu, u)), \quad \Theta_\mu(u) = \int_0^u T_\mu(\tau)\,\mathrm{d}\tau, \qquad \forall u \in \mathbb{R}. \tag{5.63}$$

We want to show that the sequence $(u_\eta^q a(u_\eta)\nabla u_\eta)_\eta$ is a Cauchy sequence in measure, this yields that, up to extract a subsequence, $u_\eta^q a(u_\eta)\nabla u_\eta \longrightarrow u^q a(u)\nabla u$ almost everywhere in $\Omega(\mathbf{x},t)$.

To do that, it suffices to prove that, for two sequences $(u_\eta)_\eta$ and $(u_{\eta'})_{\eta'}$ satisfying the saturation equation (5.14), we have

$$\mathrm{meas}\left\{ \left| u_\eta^q \nabla A(u_\eta) - u_{\eta'}^q \nabla A(u_{\eta'}) \right| \geq \delta \right\} \leq \varepsilon, \qquad \forall \varepsilon > 0. \tag{5.64}$$

We denote by $A_{\eta,\eta'} = A(u_\eta) - A(u_{\eta'})$ and $B_{\eta,\eta'} = B(u_\eta) - B(u_{\eta'})$.

We subtract the saturation equations satisfied by $(u_\eta)_\eta$ and $\left(u_{\eta'}\right)_{\eta'}$, then we multiply by $\sigma_\eta = b(u_\eta) T_\mu\left(B_{\eta,\eta'}\right)$ et $\sigma_{\eta'} = b\left(u_{\eta'}\right) T_\mu\left(B_{\eta,\eta'}\right)$ respectively, one gets

$$\begin{aligned}
&\int_\Omega \theta_\mu\left(B_{\eta,\eta'}(t,x)\right) \mathrm{d}\mathbf{x} + \int_{Q_t} \left(\nabla A(u_\eta)\cdot\nabla\sigma_\eta - \nabla A\left(u_{\eta'}\right)\cdot\nabla\sigma_{\eta'}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&\quad = \int_{Q_t} \left(\left(f(u_\eta) - g(u_\eta)\right)\mathbf{V}\cdot\nabla\sigma_\eta - \left(f\left(u_{\eta'}\right) - g\left(u_{\eta'}\right)\right)\mathbf{V}\cdot\nabla\sigma_{\eta'}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&\quad - \int_{Q_t} \left(g'(u_\eta)\nabla u_\eta\cdot\mathbf{V}\sigma_\eta - g'\left(u_{\eta'}\right)\nabla u_{\eta'}\cdot\mathbf{V}\sigma_{\eta'}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&\quad - \int_{Q_t} \eta\nabla u_\eta\cdot\nabla\sigma_\eta\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \eta'\int_{Q_t}\nabla u_{\eta'}\cdot\nabla\sigma_{\eta'}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&\quad - \int_{Q_t} \left(\nabla A(u_\eta)\cdot\mathbf{V}\sigma_\eta - \nabla A\left(u_{\eta'}\right)\cdot\mathbf{V}\sigma_{\eta'}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&\quad - \int_{Q_t} \eta\nabla u_\eta\cdot\mathbf{V}\sigma_\eta\,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \eta'\int_{Q_t}\nabla u_{\eta'}\cdot\mathbf{V}\sigma_{\eta'}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.
\end{aligned} \tag{5.65}$$

Thanks to lemma 5.13 and to the assumption (H4b), one can deduce that the sequences

$$(\nabla A(u_\eta))_\eta, (\nabla B(u_\eta))_\eta, (\nabla b(u_\eta))_\eta \text{ are uniformly bounded in } \left(L^2(Q_T)\right)^d, \tag{5.66}$$

Indeed, we have the following estimates

$$\begin{aligned}
&\|\nabla A(u_\eta)\|^2_{(L^2(Q_T))^d} = \|a(u_\eta)\nabla u_\eta\|^2_{(L^2(Q_T))^d} \le C_{M_1,m_1,r,a}, \\
&\|\nabla B(u_\eta)\|^2_{(L^2(Q_T))^d} = \|2A(u_\eta)\nabla A(u_\eta)\|^2_{(L^2(Q_T))^d} \le (2M_1)^2\|\nabla A(u_\eta)\|^2_{(L^2(Q_T))^d}.
\end{aligned}$$

It remains to show that $(\nabla b(u_\eta))_\eta$ is uniformly bounded in $\left(L^2(Q_T)\right)^d$. We have, form the definition of b, that $\nabla b(u_\eta) = 2a(u_\eta)\nabla A(u_\eta) + 2a'(u_\eta)A(u_\eta)\nabla u_\eta$. One can get the result using the following statement

$$\left|a'(u_\eta)A(u_\eta)\right| \le M_1^2 r_1 u_\eta^{r_1-1}\frac{u_\eta^{r_1+1}}{r_1+1} \le M_1^2 u_\eta^{2r_1} \le \frac{M_1^2}{m_1}u_\eta^{r_1}a(u_\eta) \le \frac{M_1^2}{m_1}a(u_\eta). \tag{5.67}$$

We denote by I_i, $i = 1, 7$, the integrals on the right-hand side of equation (5.65), and let $(\delta_\eta)_\eta$, $\left(\delta_{\eta'}\right)_{\eta'}$, $(\mathbf{V}_\eta)_\eta$, and $(\mathbf{V}_\eta)_{\eta'}$ be the sequences given by

$$\delta_\eta = \left(f(u_\eta) - g(u_\eta)\right), \quad \delta_{\eta'} = \left(f\left(u_{\eta'}\right) - g\left(u_{\eta'}\right)\right), \quad \mathbf{V}_\eta = \delta_\eta\mathbf{V}, \quad \mathbf{V}_{\eta'} = \delta_{\eta'}\mathbf{V}.$$

Using the Lebesgue theorem, we get

$$\left\|V_\eta - V_{\eta'}\right\|_{L^2(Q_T)} = \left\|\left(f(u_\eta) - g(u_\eta)\right)\mathbf{V} - \left(f\left(u_{\eta'}\right) - g\left(u_{\eta'}\right)\right)\mathbf{V}\right\|_{L^2(Q_T)} \xrightarrow[\eta,\eta'\to 0]{} 0. \tag{5.68}$$

Now, we give estimates on each term on the right-hand side of equation (5.65). For the first term, we have

$$\begin{aligned}
&\left(f(u_\eta) - g(u_\eta)\right)\mathbf{V}\cdot\nabla\sigma_\eta - \left(f\left(u_{\eta'}\right) - g\left(u_{\eta'}\right)\right)\mathbf{V}\cdot\nabla\sigma_{\eta'} \\
&= \left(\mathbf{V}_\eta\cdot\nabla b(u_\eta) - \mathbf{V}_{\eta'}\cdot\nabla b\left(u_{\eta'}\right)\right)T_\mu\left(B_{\eta,\eta'}\right) + \left(\mathbf{V}_\eta b(u_\eta) - \mathbf{V}_{\eta'}b\left(u_{\eta'}\right)\right)\nabla T_\mu\left(B_{\eta,\eta'}\right) \\
&= \left(\mathbf{V}_\eta\cdot\nabla b(u_\eta) - \mathbf{V}_{\eta'}\cdot\nabla b\left(u_{\eta'}\right)\right)T_\mu\left(B_{\eta,\eta'}\right) + \left(\mathbf{V}_\eta - \mathbf{V}_{\eta'}\right)b(u_\eta)\nabla T_\mu\left(B_{\eta,\eta'}\right) \\
&+ \left(b(u_\eta) - b\left(u_{\eta'}\right)\right)\mathbf{V}_{\eta'}\cdot\nabla T_\mu\left(B_{\eta,\eta'}\right).
\end{aligned}$$

As a consequence,

$$\begin{aligned}|I_1| \leq & \left\| \left(\mathbf{V}_\eta \cdot \nabla b\left(u_\eta\right) - \mathbf{V}_{\eta'} \cdot \nabla b\left(u_{\eta'}\right)\right) T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^1(Q_T)} \\ & + \|b\|^2_{L^\infty(Q_T)} \left\|\mathbf{V}_\eta - \mathbf{V}_{\eta'}\right\|_{(L^2(Q_T))^d} \left\|\nabla T_\mu\left(B_{\eta,\eta'}\right)\right\|_{(L^2(Q_T))^d} \\ & + \left\| \left(b\left(u_\eta\right) - b\left(u_{\eta'}\right)\right) \mathbf{V}_{\eta'} \cdot \nabla T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^1(Q_T)}.\end{aligned} \tag{5.69}$$

The first term on the right-hand side of inequality (5.69) is estimated, using (5.66), as follows

$$\begin{aligned}& \left\| \left(\mathbf{V}_\eta \cdot \nabla b\left(u_\eta\right) - \mathbf{V}_{\eta'} \cdot \nabla b\left(u_{\eta'}\right)\right) T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^1(Q_T)} \\ & \leq \left\|\mathbf{V}_\eta \cdot \nabla b\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^1(Q_T)} + \left\|\mathbf{V}_{\eta'} \cdot \nabla b\left(u_{\eta'}\right) T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^1(Q_T)} \\ & \leq C^2 \|\mathbf{V}\|^2_{(L^\infty(Q_T))^d} \left\|T_\mu\left(B_{\eta,\eta'}\right)\right\|_{(L^2(Q_T))^d} \left(\left\|\nabla b\left(u_\eta\right)\right\|_{(L^2(Q_T))^d} + \left\|\nabla b\left(u_{\eta'}\right)\right\|_{(L^2(Q_T))^d}\right),\end{aligned}$$

where C is the nonnegative constant introduced in assumption (H2) independent of η.
Taking into account the uniform boundedness in $\left(L^2\left(Q_T\right)\right)^d$ of the sequence $\left(\nabla b\left(u_\eta\right)\right)_\eta$, and the following overestimate $\left|T_\mu\left(B_{\eta,\eta'}\right)\right| \leq \mu$ as well as the convergence $B_{\eta,\eta'}$ to zero for almost every $(\mathbf{x}, t) \in Q_T$, one has

$$\left\| \left(\mathbf{V}_\eta \cdot \nabla b\left(u_\eta\right) - \mathbf{V}_{\eta'} \cdot \nabla b\left(u_{\eta'}\right)\right) T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^1(Q_T)} \xrightarrow[\eta,\eta' \to 0]{} 0.$$

It is easy to see, using the convergence (5.68), that the second term on the right-hand side of inequality (5.69) tends to zero as $\eta, \eta' \to 0$.

It remains to prove that the last term on the right-hand side of inequality (5.69) tends to zero as $\eta, \eta' \to 0$. Indeed, using the boundedness of the functions b (due to the boundedness of the dissipation function a and consequently to the function A), and thanks to the Lebsgue theorem, one has $\left\| \left(b\left(u_\eta\right) - b\left(u_{\eta'}\right)\right) \mathbf{V}_{\eta'}\right\|_{(L^2(Q_T))^d}$ tends to zero as $\eta, \eta' \to 0$. One can conclude the result, using statement (5.66) and the boundedness of the sequence $\left(\nabla T_\mu\left(B_{\eta,\eta'}\right)\right)_{\eta,\eta'}$.
For the second term of equation (5.65), we have using the definition of the function b that

$$\begin{aligned}& g'\left(u_\eta\right) \nabla u_\eta \cdot \mathbf{V} b\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) - g'\left(u_{\eta'}\right) \nabla u_{\eta'} \cdot \mathbf{V} b\left(u_{\eta'}\right) T_\mu\left(B_{\eta,\eta'}\right) \\ & \quad = \left(g'\left(u_\eta\right) \mathbf{V} \cdot \nabla B\left(u_\eta\right) - g'\left(u_{\eta'}\right) \mathbf{V} \cdot \nabla B\left(u_{\eta'}\right)\right) T_\mu\left(B_{\eta,\eta'}\right).\end{aligned}$$

The Hölder inequality and Lebesgue's theorem ensure that

$$|I_2| \leq C_{g',\mathbf{V}} \left\|T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^2(Q_T)} \left\|\nabla B\left(u_\eta\right)\right\|_{(L^2(Q_T))^d} + \left\|\nabla B\left(u_\eta\right)\right\|_{(L^2(Q_T))^d} \xrightarrow[\eta,\eta' \to 0]{} 0.$$

For the third term I_3 of equation (5.65), we write

$$\begin{aligned}\nabla u_\eta \cdot \nabla \left(b\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right)\right) = & \, 2 \nabla A\left(u_\eta\right) \cdot \nabla A\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) \\ & + 2A\left(u_\eta\right) \cdot \nabla A\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) + 2A\left(u_\eta\right) a'\left(u_\eta\right) \nabla u_\eta \cdot \nabla u_\eta T_\mu\left(B_{\eta,\eta'}\right).\end{aligned}$$

Using statements (5.67)–(5.66), one can deduce with the help of the Hölder inequality, that $|I_4| \leq \kappa\eta$, for some constant $\kappa > 0$ independent of η and η'. Therefore, $|I_4| \xrightarrow[\eta,\eta' \to 0]{} 0$.
For the fifth term I_5 of equation (5.65), we write

$$\begin{aligned}I_5 &= \int_{Q_t} \nabla A\left(u_{\eta'}\right) \cdot \mathbf{V} b\left(u_{\eta'}\right) T_\mu\left(B_{\eta,\eta'}\right) - \nabla A\left(u_\eta\right) \cdot \mathbf{V} b\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &= \int_{Q_t} \left(a\left(u_{\eta'}\right) \mathbf{V} \cdot \nabla B\left(u_{\eta'}\right) - a\left(u_\eta\right) \mathbf{V} \cdot \nabla B\left(u_\eta\right)\right) T_\mu\left(B_{\eta,\eta'}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t.\end{aligned}$$

Using again Hölder's inequality and the Lebesgue theorem, one can conclude that

$$|I_5| \leq \left\| \left(a\left(u_\eta\right) \mathbf{V} \cdot \nabla B\left(u_\eta\right) - a\left(u_{\eta'}\right) \mathbf{V} \cdot \nabla B\left(u_{\eta'}\right) \right) T_\mu\left(B_{\eta,\eta'}\right) \right\|_{L^1(Q_T)} \xrightarrow[\eta,\eta' \to 0]{} 0.$$

Finally, for the last two terms of equation (5.65), we have

$$\int_{Q_t} \nabla u_\eta \cdot \mathbf{V} b\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t = 2 \int_{Q_t} \nabla A\left(u_\eta\right) \cdot \mathbf{V} A\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}t.$$

As a consequence,

$$|I_6| \leq \eta \left(\|\nabla A\left(u_\eta\right)\|_{(L^2(Q_T))^d} \|\mathbf{V}\|_{(L^2(Q_T))^d} M_1 \mu \right) \leq \kappa \eta \xrightarrow[\eta,\eta' \to 0]{} 0.$$

Similarly, we prove that $|I_7| \leq \kappa\eta \xrightarrow[\eta,\eta' \to 0]{} 0$ for some constant $\kappa > 0$ independent of η and η'.

We denote by $W_\mu\left(\eta, \eta'\right)$ the right-hand side of equation (5.65) and by $V\left(\mu\right)$ the firm term on the left-hand side of the same equation ; from the estimations on the integrals I_i, $W_\mu\left(\eta, \eta'\right)$ goes to zero as $\eta, \eta' \to 0$, for all $\mu > 0$. We also have $|V\left(\mu\right)| \leq |\Omega|\, \mu$, which goes to zero as $\mu \to 0$ and uniformly on η and η'. Therefore, we have the following result stemming from equation (5.65) and the aforementioned definitions

$$\begin{aligned} \int_{Q_T} \left(\nabla A\left(u_\eta\right) \cdot \nabla \left(b\left(u_\eta\right) T_\mu\left(B_{\eta,\eta'}\right) \right) - \nabla A\left(u_{\eta'}\right) \cdot \nabla \left(b\left(u_{\eta'}\right) T_\mu\left(B_{\eta,\eta'}\right) \right) \right) \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ = W_\mu\left(\eta, \eta'\right) + V\left(\mu\right). \end{aligned} \tag{5.70}$$

Here, we give the second step for the proof of statement (5.64). Let s the continuous function defined by

$$s\left(u\right) = \int_0^u b\left(z\right) A\left(z\right) a\left(z\right)\, \mathrm{d}z, \qquad \forall u \in \mathbb{R}. \tag{5.71}$$

Let us prove that

$$\mathrm{meas}\left\{ \left| \nabla s\left(u_\eta\right) - \nabla s\left(u_{\eta'}\right) \right| \geq \delta \right\} \xrightarrow[\eta,\eta' \to 0]{} 0.$$

Remark that $\left\{ \left| \nabla s\left(u_\eta\right) - \nabla s\left(u_{\eta'}\right) \right| \geq \delta \right\} \subset \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4$, where

$$\begin{aligned} &\mathcal{A}_1 = \left\{ |\nabla A\left(u_\eta\right)| \geq k \right\}, \quad \mathcal{A}_2 = \left\{ \left| \nabla A\left(u_{\eta'}\right) \right| \geq k \right\}, \quad \mathcal{A}_3 = \left\{ \left| B_{\eta,\eta'} \right| \geq k \right\}, \\ &\mathcal{A}_4 = \left\{ \left| \nabla s\left(u_\eta\right) - \nabla s\left(u_{\eta'}\right) \right| \geq \delta \right\} \cap \left\{ |\nabla A\left(u_\eta\right)| \leq k \right\} \cap \left\{ \left| \nabla A\left(u_{\eta'}\right) \right| \leq k \right\} \cap \left\{ \left| B_{\eta,\eta'} \right| \leq \mu \right\}. \end{aligned}$$

Thanks to the statement (5.66), and the continuous embedding of $L^2\left(Q_T\right)$ in $L^1\left(Q_T\right)$, we have

$$\begin{aligned} \mathrm{meas}\left(\mathcal{A}_1\right) k \leq \int_{\mathcal{A}_1} |\nabla A\left(u_\eta\right)|\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \leq \int_{Q_T} |\nabla A\left(u_\eta\right)|\, \mathrm{d}\mathbf{x}\, \mathrm{d}t = \|\nabla A\left(u_\eta\right)\|_{(L^1(Q_T))^d} \\ \leq C_\Omega \|\nabla A\left(u_\eta\right)\|_{(L^2(Q_T))^d} \leq C_{M_1, m_1, r, a, \Omega}. \end{aligned}$$

An analogous estimate hols for $\mathcal{A}_2$. Therefore, by choosing k large enough, one gets meas $\left(\mathcal{A}_1\right)$ + meas $\left(\mathcal{A}_2\right)$ is arbitrarily small. In the same manner, one gets

$$\mathrm{mes}\left(\mathcal{A}_3\right) \leq \frac{1}{\mu} \left\| B_{\eta,\eta'} \right\|_{L^1(Q_T)},$$

which, for a fixed $\mu > 0$, tends to zero as $\eta, \eta' \to 0$.
It remains to show that meas$(\mathcal{A}_4)$ is small enough. Indeed, we have

$$\begin{aligned} \left| \nabla s\left(u_\eta\right) - \nabla s\left(u_{\eta'}\right) \right|^2 &= \left| b\left(u_\eta\right) A\left(u_\eta\right) \nabla A\left(u_\eta\right) - b\left(u_{\eta'}\right) A\left(u_{\eta'}\right) \nabla A\left(u_{\eta'}\right) \right|^2 \\ &= \left| b\left(u_\eta\right) A\left(u_\eta\right) \nabla A_{\eta,\eta'} + \left(b\left(u_\eta\right) A\left(u_\eta\right) - b\left(u_{\eta'}\right) A\left(u_{\eta'}\right) \right) \nabla A\left(u_{\eta'}\right) \right|^2, \end{aligned}$$

and therefore, one gets

$$\begin{aligned}\delta\text{meas}\left(\mathcal{A}_4\right) &\leq \int_{\mathcal{A}_4}\left|\nabla s\left(u_\eta\right)-\nabla s\left(u_{\eta'}\right)\right|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq 2\int_{\mathcal{A}_4}\left|b\left(u_\eta\right)A\left(u_\eta\right)\nabla A_{\eta,\eta'}\right|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &+2\int_{\mathcal{A}_4}\left|b\left(u_\eta\right)A\left(u_\eta\right)-b\left(u_{\eta'}\right)A\left(u_{\eta'}\right)\right|^2\left|\nabla A\left(u_{\eta'}\right)\right|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &\leq 4M_1^3\int_{\mathcal{A}_4}b\left(u_\eta\right)\left(A\left(u_\eta\right)+A\left(u_{\eta'}\right)\right)\nabla A_{\eta,\eta'}\cdot\nabla A_{\eta,\eta'}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &+2k^2\int_{Q_T}\left|b\left(u_\eta\right)A\left(u_\eta\right)-b\left(u_{\eta'}\right)A\left(u_{\eta'}\right)\right|^2\mathrm{d}\mathbf{x}\,\mathrm{d}t\end{aligned}$$

The parameter k is chosen to be fixed and large enough; then the last term that we denote $W_k\left(\eta,\eta'\right)$ goes to zero as $\eta,\eta'\to 0$. Consequently,

$$\begin{aligned}\delta\text{meas}\left(\mathcal{A}_4\right) &\leq W_k\left(\eta,\eta'\right)\\ &+4M_1^3\int_{Q_T}b\left(u_\eta\right)\left(A\left(u_\eta\right)+A\left(u_{\eta'}\right)\right)\nabla A_{\eta,\eta'}\cdot\nabla A_{\eta,\eta'}1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\end{aligned}\tag{5.72}$$

We want to show that the last term on the left-hand side of inequality (5.72) is small enough. For that, we compute

$$\begin{aligned}\nabla A_{\eta,\eta'}\cdot\nabla\sigma_\eta &= b\left(u_\eta\right)\nabla A_{\eta,\eta'}\cdot\nabla T_\mu\left(B_{\eta,\eta'}\right)+\nabla A_{\eta,\eta'}\cdot\nabla b\left(u_{\eta'}\right)T_\mu\left(B_{\eta,\eta'}\right)\\ &= b\left(u_\eta\right)\nabla A_{\eta,\eta'}\cdot\nabla B_{\eta,\eta'}1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}+\nabla A_{\eta,\eta'}\cdot\nabla b\left(u_{\eta'}\right)T_\mu\left(B_{\eta,\eta'}\right)\\ &= 2b\left(u_\eta\right)\nabla A_{\eta,\eta'}\cdot\nabla A\left(u_\eta\right)A\left(u_\eta\right)1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}\\ &\quad-2b\left(u_\eta\right)\nabla A_{\eta,\eta'}\cdot\nabla A\left(u_\eta\right)A\left(u_{\eta'}\right)1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}+\nabla A_{\eta,\eta'}\cdot\nabla b\left(u_{\eta'}\right)T_\mu\left(B_{\eta,\eta'}\right)\\ &= b\left(u_\eta\right)\left(A\left(u_\eta\right)+A\left(u_{\eta'}\right)\right)\nabla A_{\eta,\eta'}\cdot\nabla A_{\eta,\eta'}1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}\\ &\quad+\left(b\left(u_\eta\right)A_{\eta,\eta'}\nabla\left(A\left(u_\eta\right)+A\left(u_{\eta'}\right)\right)1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}+\nabla b\left(u_{\eta'}\right)T_\mu\left(B_{\eta,\eta'}\right)\right)\cdot\nabla A_{\eta,\eta'}.\end{aligned}\tag{5.73}$$

We have the following estimates

$$\begin{aligned}&\left\|\nabla b\left(u_{\eta'}\right)T_\mu\left(B_{\eta,\eta'}\right)\cdot\nabla A_{\eta,\eta'}\right\|_{(L^1(Q_T))^d}\\ &\qquad\leq\left\|T_\mu\left(B_{\eta,\eta'}\right)\right\|_{L^\infty(Q_T)}\left\|\nabla A_{\eta,\eta'}\right\|_{(L^2(Q_T))^d}\left\|\nabla b\left(u_{\eta'}\right)\right\|_{(L^2(Q_T))^d}\leq\kappa\mu,\end{aligned}$$

and

$$\begin{aligned}&\left\|b\left(u_\eta\right)\nabla A_{\eta,\eta'}\cdot\nabla\left(A\left(u_\eta\right)+A\left(u_{\eta'}\right)\right)A_{\eta,\eta'}1_{\left\{|B_{\eta,\eta'}|\leq\mu\right\}}\right\|_{(L^1(Q_T))^d}\\ &\qquad\leq 2\mu M_1^2\left\|\nabla A_{\eta,\eta'}\right\|_{(L^2(Q_T))^d}\left\|\nabla\left(A\left(u_\eta\right)+A\left(u_{\eta'}\right)\right)\right\|_{(L^2(Q_T))^d}\leq\kappa\mu.\end{aligned}$$

Consequently, the two last terms of equation (5.73) tend to zero in $L^1\left(Q_T\right)$ as μ goes to zero and therefore, they can be included into the function $V\left(\mu\right)$. Now, we are interested in giving an

estimate on the left-hand side on equation (5.73). Indeed, we have

$$\begin{aligned}\int_{Q_t}\nabla A_{\eta,\eta'}\cdot\nabla\sigma_\eta\,\mathrm{d}\mathbf{x}\,\mathrm{d}t &= \int_{Q_t}\nabla A_{\eta,\eta'}\cdot\nabla\left(b\left(u_\eta\right)T_\mu\left(B_{\eta,\eta'}\right)\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &= \int_{Q_T}\left(\nabla A\left(u_\eta\right)\cdot\nabla\left(b\left(u_\eta\right)T_\mu\left(B_{\eta,\eta'}\right)\right)-\nabla A\left(u_{\eta'}\right)\cdot\nabla\left(b\left(u_{\eta'}\right)T_\mu\left(B_{\eta,\eta'}\right)\right)\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &- \int_{Q_T}\nabla A\left(u_{\eta'}\right)\cdot\nabla\left(b\left(u_\eta\right)-b\left(u_{\eta'}\right)\right)T_\mu\left(B_{\eta,\eta'}\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &- \int_{Q_T}\left(b\left(u_\eta\right)-b\left(u_{\eta'}\right)\right)\nabla A\left(u_{\eta'}\right)\cdot\nabla B_{\eta,\eta'}1_{\left\{|B_{\eta,\eta'}|\le\mu\right\}}\mathrm{d}\mathbf{x}\,\mathrm{d}t.\end{aligned}$$

It is easy to see, using the previous results, that

$$\int_{Q_t}\nabla A_{\eta,\eta'}\cdot\nabla\sigma_\eta\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\le W_\mu\left(\eta,\eta'\right)+\kappa\mu \tag{5.74}$$

Finally, using estimates (5.73)–(5.74), then inequality (5.72) gives

$$\delta\mathrm{mes}\left(\mathcal{A}_4\right)\le W_\mu\left(\eta,\eta'\right)+V\left(\mu\right)+W_k\left(\eta,\eta'\right)$$

Using the above results, one can deduce that for all $\varepsilon>0$, for all $\delta>0$, there exists $\eta_0>0$ such that for all $\eta,\eta'\le\eta_0$, we have

$$\mathrm{meas}\left\{\left|\nabla s\left(u_\eta\right)-\nabla s\left(u_{\eta'}\right)\right|\ge\delta\right\}\le\varepsilon \tag{5.75}$$

Now, we can prove statement (5.64) with the help of inequality (5.75). Indeed, we have

$$u_\eta^q\nabla A\left(u_\eta\right)-u_{\eta'}^q\nabla A\left(u_{\eta'}\right)=u_\eta^q\nabla A_{\eta,\eta'}+\left(u_\eta^q-u_{\eta'}^q\right)\nabla A\left(u_{\eta'}\right).$$

Since $q=3r_1+2$, then $u_\eta^q=u_\eta^{3r_1+2}\le C_{r_1,m_1}b\left(u_\eta\right)A\left(u_\eta\right)$ where $C_{r_1,m_1}=\dfrac{\left(r_1+1\right)^2}{2m_1^3}$.

We write

$$\begin{aligned}\left|u_\eta^q\nabla A_{\eta,\eta'}\right|\le&C_{r_1,m_1}b\left(u_\eta\right)A\left(u_\eta\right)\left|\nabla A_{\eta,\eta'}\right|\\ &\le C_{r_1,m_1}\left|\nabla s\left(u_\eta\right)-\nabla s\left(u_{\eta'}\right)\right|+C_{r_1,m_1}\left|\left(b\left(u_\eta\right)A\left(u_\eta\right)-b\left(u_{\eta'}\right)A\left(u_{\eta'}\right)\right)\nabla A\left(u_{\eta'}\right)\right|.\end{aligned}$$

Consequently,

$$\begin{aligned}\left|u_\eta^q\nabla A\left(u_\eta\right)-u_{\eta'}^q\nabla A\left(u_{\eta'}\right)\right|\le&\left|\left(u_{\eta'}^q-u_\eta^q\right)\nabla A\left(u_{\eta'}\right)\right|+C_{r_1,m_1}\left|\nabla s\left(u_\eta\right)-\nabla s\left(u_{\eta'}\right)\right|\\ &+C_{r_1,m_1}\left|\left(b\left(u_\eta\right)A\left(u_\eta\right)-b\left(u_{\eta'}\right)A\left(u_{\eta'}\right)\right)\nabla A\left(u_{\eta'}\right)\right|,\end{aligned}$$

which converges to zero as $\eta,\eta'\to0$. The result is due either to the convergence in $L^1\left(Q_T\right)$ for the first and the last terms on the right-hand side, or either by the help of (5.75). This ends the proof of lemma 5.14. □

Thanks to lemma 5.13 and lemma 5.14, we have the following lemma

Lemma 5.15. *Let $q=3r_1+2$, we have the following convergences as η goes to zero*

$$u_\eta\longrightarrow u\text{ strongly in }L^2(Q_T), \tag{5.76}$$

$$h_\theta\left(u_\eta\right)\longrightarrow h_\theta\left(u\right)\text{ strongly in }L^2\left(Q_T\right), \tag{5.77}$$

$$\sqrt{u_\eta^{r-2}a\left(u_\eta\right)}\nabla u_\eta\longrightarrow\sqrt{u^{r-2}a\left(u\right)}\nabla u\text{ weakly in }\left(L^2\left(Q_T\right)\right)^d, \tag{5.78}$$

$$u_\eta^{r-1}\nabla u_\eta\longrightarrow u^{r-1}\nabla u\text{ weakly in }\left(L^2\left(Q_T\right)\right)^d, \tag{5.79}$$

Furthermore,

$$u_\eta \longrightarrow u \text{ almost everywhere in } Q_T, \tag{5.80}$$
$$0 \leq u(\mathbf{x},t) \leq 1 \text{ almost everywhere in } Q_T, \tag{5.81}$$
$$u_\eta^q a(u_\eta)\nabla u_\eta \longrightarrow u^q a(u)\nabla u \text{ almost everywhere in } \left(L^2(Q_T)\right)^d, \tag{5.82}$$

Proof. To prove the strong convergence (5.76), we deduce from lemma (5.13), thanks to the Aubin–Lions lemma, that there exists a subsequence of $(u_\eta)_\eta$ such that

$$u_\eta \longrightarrow u \text{ strongly in } L^2(Q_T).$$

As a consequence, $h_\theta(u_\eta) \longrightarrow h_\theta(u)$ as $\eta \to 0$ strongly in $L^2(Q_T)$ since h_θ is a continuous function from $L^2(Q_T)$ to $L^2(Q_T)$.
To prove statement (5.78), we consider the continuous function $\gamma : L^2(Q_T) \longrightarrow L^2(Q_T)$ defined by $\gamma(u_\eta) = \int_0^{u_\eta} \sqrt{\tau^{r-2}a(\tau)}\,d\tau$. We have, $\gamma(u_\eta) \xrightarrow[\eta\to 0]{} \gamma(u)$ fortement dans $L^2(Q_T)$, and since the sequence $(\nabla\gamma(u_\eta))_\eta = \left(\sqrt{u_\eta a(u_\eta)}\nabla u_\eta\right)_\eta$ converges weakly in $\left(L^2(Q_T)\right)^d$, then it is easy to see that

$$\nabla\gamma(u_\eta) = u_\eta^q a(u_\eta)\nabla u_\eta \xrightarrow[\eta\to 0]{} \nabla\gamma(u) = u^q a(u)\nabla u, \text{ weakly in } \left(L^2(Q_T)\right)^d.$$

In the same manner, we prove the convergence result (5.79). The convergence (5.80) is a consequence of the strong convergence (5.76), while the statement (5.81) is obtained using the fact that $\{u \in L^2(Q_T),\ 0 \leq u(\mathbf{x},t) \leq 1 \text{ for a.e. } (\mathbf{x},t) \in Q_T\}$ is a closed subset of $L^2(Q_T)$.
Finally, the proof of the almost everywhere convergence (5.82) is a direct consequence of lemma 5.14. □

To complete the proof of theorem 5.4, we pass to the limit as η goes to zero in the following weak formulation : $\forall\chi \in \mathcal{C}^1\left([0,T];H_0^1(\Omega)\right)$ with $\chi(\cdot,T)=0$,

$$\begin{aligned}
&-\int_0^T h_\theta(u_\eta)\partial_t\chi\,d\mathbf{x}\,dt - \int_\Omega h_\theta(u_0)\chi(\mathbf{x},0)\,d\mathbf{x} + \int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\chi\,d\mathbf{x}\,dt\\
&+\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\nabla\chi\,\beta_\theta(u_\eta)\,d\mathbf{x}\,dt + \eta\int_{Q_T}\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\chi\,d\mathbf{x}\,dt\\
&+\eta\int_{Q_T}\nabla u_\eta\cdot\nabla\chi\,\beta_\theta(u_\eta)\,d\mathbf{x}\,dt - \int_{Q_T}(f(u_\eta)-g(u_\eta))\,\mathbf{V}\cdot\nabla\beta_\theta(u_\eta)\chi\,d\mathbf{x}\,dt\\
&+\int_{Q_T} g'(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\beta_\theta(u_\eta)\chi\,d\mathbf{x}\,dt - \int_{Q_T}(f(u_\eta)-g(u_\eta))\,\mathbf{V}\cdot\nabla\chi\,\beta_\theta(u_\eta)\,d\mathbf{x}\,dt\\
&+\int_{Q_T} a(u_\eta)\,\mathbf{V}\cdot\nabla u_\eta\,\beta_\theta(u_\eta)\chi\,d\mathbf{x}\,dt + \eta\int_{Q_T}\nabla u_\eta\cdot\mathbf{V}\beta_\theta(u_\eta)\chi\,d\mathbf{x}\,dt = 0.
\end{aligned} \tag{5.83}$$

Indeed, we denote by L_i, $i=1,...,11$ the integral terms in equation (5.83).
From the definition of the continuous function β_θ, we have

$$\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\nabla\chi\beta_\theta(u_\eta)\,d\mathbf{x}\,dt = \int_{Q_T} u_\eta^{r-1}\nabla u_\eta\cdot u_\eta^\theta a(u_\eta)\nabla\chi\,d\mathbf{x}\,dt.$$

The sequence $\left(u_\eta^{r-1}\nabla u_\eta\right)_\eta$ converges weakly towards $u^{r-1}\nabla u$ in $\left(L^2(Q_T)\right)^d$. Further, thanks to Lebesgue's theorem, the sequence $\left(u_\eta^\theta a(u_\eta)\nabla\chi\right)_\eta$ converges strongly towards $u^\theta a(u)\nabla\chi$ in

$\left(L^2\left(Q_T\right)\right)^d$; this gives the convergences of terms L_4, and L_{10}. In the same manner, we obtain the convergence of $L_8 + L_9$ towards

$$\int_{Q_T} g'(u)\,\mathbf{V}\cdot\nabla u\,\beta_\theta(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{Q_T}\left(f(u)-g(u)\right)\mathbf{V}\cdot\nabla\chi\,\beta_\theta(u)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Let us focus on the seventh term L_7 of equation (5.83). Since $\theta > 1$, then there exists $\theta_0 > 0$ such that $\theta = 1 + \theta_0$. Therefore, using Lebesgue's theorem and the weak convergence (5.79), one has

$$\begin{aligned} L_7 &= -(r-1+\theta)\int_{Q_T} u_\eta^{r-1}\nabla u_\eta\cdot\mathbf{V}u_\eta^{\theta_0}\left(f(u_\eta)-g(u_\eta)\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\xrightarrow[\eta\to 0]{} \int_{Q_T}\left(f(u)-g(u)\right)\mathbf{V}\cdot\nabla\beta_\theta(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \end{aligned}$$

For the fifth term, we have

$$\begin{aligned} |L_5| &= \eta\left|\int_{Q_T}\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right| = \kappa\eta\int_{Q_T}\left|u_\eta^{r-2}u_\eta^{\theta}\chi\nabla u_\eta\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\le \kappa\eta\int_{Q_T}\left|u_\eta^{2r-2}\chi\nabla u_\eta\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t \le \kappa\eta\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\chi\|_{L^\infty(Q_T)}. \end{aligned}$$

As a consequence, $|L_5| \le \kappa\eta \longrightarrow 0$, as η goes to zero.
The convergence to zero for the sixth and the last terms, is similar to the that for L_7. Indeed, we have

$$\begin{aligned} |L_6| &= \eta\left|\int_{Q_T}\nabla u_\eta\cdot\nabla\chi\beta_\theta(u_\eta)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right| = \eta\left|\int_{Q_T} u_\eta^{r-1}\nabla u_\eta\cdot u_\eta^{\theta}\nabla\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right| \\ &\le \kappa\eta\left\|u_\eta^{r-1}\nabla u_\eta\right\|_{(L^2(Q_T))^d}\|\nabla\chi\|_{(L^2(Q_T))^d} \xrightarrow[\eta\to 0]{}. \end{aligned}$$

Now, we are interested in the third term L_3 of equation (5.83). One can remark that the sequence $\left(a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\right)_\eta \subset \mathbb{R}$ is nonnegative and

$$a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta) = (r-1+\theta)\,u_\eta^{r-2+\theta}a(u_\eta)\nabla u_\eta\cdot\nabla u_\eta$$

converges almost everywhere, up to a subsequence, to $a(u)\nabla u\cdot\nabla\beta_\theta(u)$, since $r-2+\theta-2q-r_1 \ge 0$, i.e. $\theta \ge 7r_1 + 6 - r$.
Consider a nonnegative test function $(\chi \ge 0)$; then the lemma of Fatou ensures that

$$\liminf_{\eta\to 0}\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \ge \int_{Q_T} a(u)\nabla u\cdot\nabla\beta_\theta(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

then the limit solution u verifies inequality (5.6) given into definition 5.3. Finally, to obtain (5.7), we apply the Egorov theorem on the sequence $\left(a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\right)_\eta$ which converges almost everywhere. Indeed, we have

$$\begin{aligned} &\forall\varepsilon>0, \exists\, Q^\varepsilon \subset Q_T \text{ such that meas}\left(Q^\varepsilon\right) < \varepsilon, \text{ and} \\ &a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta) \xrightarrow[\eta\to 0]{} a(u)\nabla u\cdot\nabla\beta_\theta(u) \text{ uniformly in } Q_T\backslash Q^\varepsilon. \end{aligned}$$

Now, we take a test function χ such that $\operatorname{supp}\chi \subset ([0,T)\times\Omega)\backslash Q^\varepsilon$, then

$$\int_{Q_T\backslash Q^\varepsilon} a(u_\eta)\nabla u_\eta\cdot\nabla\beta_\theta(u_\eta)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \xrightarrow[\eta\to 0]{} \int_{Q_T\backslash Q^\varepsilon} a(u)\nabla u\cdot\nabla\beta_\theta(u)\,\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

The proof of theorem 5.4 is now accomplished. □

5.6 Proof of theorem 5.6

The proof of theorem 5.6 is a combination of the two techniques used for theorem 5.2 and theorem 5.4.

Lemma 5.16. *Under the assumptions* $(H1)-(H3)$ *and* (H4c), *assume that* $G(u_0)$ *belongs to* $L^1(\Omega)$. *Then the solutions to the saturation equation* (5.14) *verify*

(i) $0 \leq u_\eta(\mathbf{x},t) \leq 1$, *for almost everywhere* $(\mathbf{x},t) \in Q_T$.

(ii) The sequences $\left(\sqrt{\mu'(u_\eta)\,a(u_\eta)}\nabla u_\eta\right)_\eta$ *and* $\left(a(u_\eta)\nabla u_\eta\right)_\eta$ *are uniformly bounded in* $\left(L^2(Q_T)\right)^d$.

(iii) The sequences $\left(\sqrt{\eta\mu'(u_\eta)}\nabla u_\eta\right)_\eta$ *and* $\left(\nabla J(u_\eta)\right)_\eta$ *are uniformly bounded in* $\left(L^2(Q_T)\right)^d$.

(iv) The sequence $\left(G(u_\eta)\right)_\eta$ *is uniformly bounded in* $L^\infty\left(0,T;L^2(\Omega)\right)$.

(v) The sequence $\left(\partial_t J(u_\eta)\right)_\eta$ *is uniformly bounded in* $L^1\left(0,T;W^{-1,q'}(\Omega)\right)$.

(vi) The sequences $\left(J(u_\eta)\right)_\eta$ *and* $(u_\eta)_\eta$ *are relatively compact in* $L^2\left(0,T;L^2(\Omega)\right)$.

Proof. The first part, (i), is obtained in section 5.3.2.
Now, we multiply the saturation equation (5.14) by $\mu(u_\eta)$ and integrate over Ω, this yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega G(u_\eta)\,\mathrm{d}\mathbf{x} &+ \int_\Omega a(u_\eta)\,\mu'(u_\eta)\,|\nabla u_\eta|^2\mathrm{d}\mathbf{x} + \eta\int_\Omega \mu'(u_\eta)\,|\nabla u_\eta|^2\mathrm{d}\mathbf{x}\\ &= \int_\Omega \left(f(u_\eta)-g(u_\eta)\right)\mu'(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x} - \int_\Omega g'(u_\eta)\,\mu(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\\ &- \int_\Omega a(u_\eta)\,\mu(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x} - \eta\int_\Omega \mu(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}.\end{aligned} \tag{5.84}$$

We split the whole integral appearing in equation (5.84) into two parts, similarly as we do in section 5.4, we write $\int_\Omega = \int_{\Omega\cap\{u_\eta<u_*\}} + \int_{\Omega\cap\{u_\eta\geq u_*\}}$ and we denote $\Omega_1 = \Omega\cap\{u_\eta<u_*\}$ and $\Omega_2 = \Omega\cap\{u_\eta\geq u_*\}$.

• Into the region Ω_1 ; and following the same analysis as the one developed into section 5.4, we obtain the following estimates

$$\begin{aligned}&\left|\int_{\Omega_1}\left(f(u_\eta)-g(u_\eta)\right)\mu'(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x} - \int_{\Omega_1} g'(u_\eta)\,\mu(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\right|\\ &\qquad\leq \frac{1}{4}\left\|\sqrt{\mu'(u_\eta)\,a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega_1))^d} + 2\frac{1+(C(r-1))^2}{m_1(r-1)}\|\mathbf{V}\|^2_{(L^2(\Omega_1))^d},\\ &\left|\int_{\Omega_1} a(u_\eta)\,\mu(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\right| \leq \frac{1}{4}\left\|\sqrt{\mu'(u_\eta)\,a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega_1))^d} + \frac{\|a\|^2_\infty}{r-1}\|\mathbf{V}\|^2_{(L^2(\Omega_1))^d},\\ &\left|\eta\int_{\Omega_1}\mu(u_\eta)\,\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\right| \leq \frac{1}{2}\left\|\sqrt{\eta\mu'(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(\Omega_1))^d} + \frac{1}{2(r-1)}\|\mathbf{V}\|^2_{(L^2(\Omega_1))^d}.\end{aligned}$$

• Into region Ω_2, we have

$$\int_{\Omega_2}\left(\left(f(u_\eta)-g(u_\eta)\right)\mu'(u_\eta) - g'(u_\eta)\,\mu(u_\eta)\right)\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x} = 0.$$

Furthermore, we have the following estimates

$$\begin{aligned}
&\left|\int_{\Omega_2} a\left(u_\eta\right)\mu\left(u_\eta\right)\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\right| \\
&\qquad\leq \frac{1}{2}\left\|\sqrt{\mu'\left(u_\eta\right)a\left(u_\eta\right)}\nabla u_\eta\right\|^2_{(L^2(\Omega_2))^d}+\frac{k_1\left\|\sqrt{\mu}\mu' a\right\|^2_\infty}{4}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega_2))^d}, \\
&\left|\eta\int_{\Omega_2}\mu\left(u_\eta\right)\nabla u_\eta\cdot\mathbf{V}\mathrm{d}\mathbf{x}\right|\leq\frac{1}{2}\left\|\sqrt{\eta\mu'\left(u_\eta\right)}\nabla u_\eta\right\|^2_{(L^2(\Omega_2))^d}+\frac{k_1\left\|\sqrt{\mu}\mu'\right\|^2_\infty}{4}\left\|\mathbf{V}\right\|^2_{(L^2(\Omega_2))^d}.
\end{aligned}$$

Plugging the previous estimates into equation (5.84), one has

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega G\left(u_\eta\right)\mathrm{d}\mathbf{x}+\left\|\sqrt{\mu'\left(u_\eta\right)a\left(u_\eta\right)}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d}+\left\|\sqrt{\eta\mu'\left(u_\eta\right)}\nabla u_\eta\right\|^2_{(L^2(\Omega))^d} \\
\leq\kappa\left\|\mathbf{V}\right\|^2_{(L^2(\Omega))^d}
\end{aligned}\tag{5.85}$$

Now, we integrate inequality (5.85) with respect to the time over $(0,t), t\in(0,T)$, one deduces that the sequences $\left(\sqrt{\mu'\left(u_\eta\right)a\left(u_\eta\right)}\nabla u_\eta\right)_\eta$ and $\left(\sqrt{\eta\mu'\left(u_\eta\right)}\nabla u_\eta\right)_\eta$ are uniformly bounded in $\left(L^2\left(Q_T\right)\right)^d$, and that $\left(G\left(u_\eta\right)\right)_\eta$ is uniformly bounded in $L^\infty\left(0,T;L^2\left(\Omega\right)\right)$.
Let us prove that $\left(a\left(u_\eta\right)\nabla u_\eta\right)_\eta$ and $\left(\nabla J\left(u_\eta\right)\right)_\eta$ are uniformly bounded in $\left(L^2\left(Q_T\right)\right)^d$. Indeed, for all $0\leq u\leq u_*$, we have

$$a\left(u\right)\mu'\left(u\right)\geq m_1\left(r-1\right)u^{r_1}u^{r-2}\geq m_1\left(r-1\right)u^r u^{r-2}\geq m_1\left(r-1\right)u^{2r-2}\geq c_1 j^2\left(u\right),$$

and for all $u\geq u_*$, we have

$$\begin{aligned}
a\left(u\right)\mu'\left(u\right)&\geq m_1(1-u)^{r_2}\mu\left(u\right)g'\left(u\right)\left(f\left(u\right)-g\left(u\right)\right)^{-1}\geq c_1(1-u)^{r_2-2} \\
&\geq\frac{c_1}{\left(1-u_*\right)^{2-r'}}\left(\left(1-u_*\right)^{1-\frac{r'}{2}}\right)^2\left(\left(1-u\right)^{\frac{r'}{2}-1}\right)^2\geq\kappa j^2\left(u\right),
\end{aligned}$$

Therefore, $\left\|\nabla J\left(u_\eta\right)\right\|^2_{(L^2(Q_T))^d}\leq\kappa\left\|\mu'^{\frac{1}{2}}\left(u_\eta\right)a^{\frac{1}{2}}\left(u_\eta\right)\nabla u_\eta\right\|^2_{(L^2(Q_T))^d}\leq\kappa$.
For the sequence $\left(a\left(u_\eta\right)\nabla u_\eta\right)_\eta$. It is easy to see that

$$\begin{aligned}
&a\left(u_\eta\right)\leq M_1 u_\eta^{r_1}\leq M_1 u_\eta^{r-2}\leq\frac{M_1}{r-1}\mu'\left(u_\eta\right), && \text{if } 0\leq u_\eta\leq u_*, \\
&a\left(u_\eta\right)\leq M_1(1-u_\eta)^{r_2}\leq\left(1-u_\eta\right)^{-2}\leq\kappa\mu'\left(u_\eta\right), && \text{if } u_*\leq u_\eta\leq 1.
\end{aligned}$$

As a consequence, the sequence $\left(a\left(u_\eta\right)\nabla u_\eta\right)_\eta$ is uniformly bounded in $\left(L^2\left(Q_T\right)\right)^d$.

Let us now focus on the fourth part (iv), we want to prove that

$$\left(\partial_t J\left(u_\eta\right)\right)_\eta \text{ is uniformly bounded in } L^2\left(0,T;\left(H^1\left(\Omega\right)\right)'\right)+L^1\left(Q_T\right).$$

We take a test function $\chi\in L^2\left(0,T;H^1_0\left(\Omega\right)\right)\cap L^\infty\left(Q_T\right)$ and multiply the saturation equation

(5.14) by $j(u_\eta)\chi$, this yields

$$\begin{aligned}\langle \partial_t J(u_\eta), \chi\rangle = &-\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\nabla\left(j(u_\eta)\chi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ -\eta\int_{Q_T}\nabla u_\eta\cdot\nabla\left(j(u_\eta)\chi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t&+\int_{Q_T}\left(f(u_\eta)-g(u_\eta)\right)\mathbf{V}\cdot\nabla\left(j(u_\eta)\chi\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ -\int_{Q_T} g'(u_\eta)\nabla u_\eta\cdot\mathbf{V}j(u_\eta)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t&-\int_{Q_T} a(u_\eta)\nabla u_\eta\cdot\mathbf{V}j(u_\eta)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ -\eta\int_{Q_T}\nabla u_\eta\cdot\mathbf{V}j(u_\eta)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.&\end{aligned}\tag{5.86}$$

We will give estimates on each integral on the right-hand side of equation (5.86).
Into region $Q_T\cap\{u_\eta<u_*\}$, we have $j(u_\eta)=\mu(u_\eta)$, thus we obtain the same estimates on the integrals obtained within the proof of theorem 5.4. It remains to estimate the terms of the form $\int_{Q_T\cap\{u_\eta\geq u_*\}}$ that we denote them by $\{L_i\}_{1\leq i\leq 6}$ respectively.
For the first term L_1, we have

$$\begin{aligned}|L_1|&\leq\int_{Q_T\cap\{u_\eta\geq u_*\}}\left|a(u_\eta)j'(u_\eta)\nabla u_\eta\cdot\nabla u_\eta\chi\right|+\left|a(u_\eta)j(u_\eta)\nabla u_\eta\cdot\nabla\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &\leq\left\|\sqrt{j'(u_\eta)a(u_\eta)}\nabla u_\eta\right\|^2_{(L^2(Q_T))^d}\|\chi\|_{L^\infty(Q_T)}+\|a(u_\eta)\nabla J(u_\eta)\|_{(L^2(Q_T))^d}\|\nabla\chi\|_{(L^2(Q_T))^d}.\end{aligned}$$

On the other hand, using the definition of μ and j, we have, for all $u_*\leq u\leq 1$, that

$$\left|j'(u_\eta)\right|=\left(\frac{r'}{2}-1\right)\beta(u_*)(1-u_*)^{1-\frac{r'}{2}}(1-u_\eta)^{\frac{r'}{2}-2}\leq C_{u_*,r'}(1-u_\eta)^{-2}\leq\mu'(u_\eta),$$

thus, thanks to parts (i) and (ii), one deduces that $|L_1|\leq\kappa\left(\|\chi\|_{L^\infty(Q_T)}+\|\nabla\chi\|_{(L^2(Q_T))^d}\right)$.
In the same manner, we obtain the estimates on the remaining terms except the estimate on L_3. Indeed, using the assumption (H2) on f and g, one has

$$f(u_\eta)-g(u_\eta)\leq C_{g'}(1-u_\eta),\qquad\forall u_*\leq u_\eta\leq 1,$$

and therefore, we obtain the following estimates

$$\begin{aligned}\left|\int_{Q_T}\left(f(u_\eta)-g(u_\eta)\right)j'(u_\eta)\nabla u_\eta\cdot\mathbf{V}\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right|&\leq C_{g'}\int_{Q_T}\left|j(u_\eta)\nabla u_\eta\cdot\mathbf{V}\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &\leq C_{g'}\|\nabla J(u_\eta)\|_{(L^2(Q_T))^d}\|\mathbf{V}\|_{(L^2(Q_T))^d}\|\chi\|_{L^\infty(Q_T)}.\\ \left|\int_{Q_T}\left(f(u_\eta)-g(u_\eta)\right)j(u_\eta)\mathbf{V}\cdot\nabla\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right|&\leq C_j\int_{Q_T}\left|\mathbf{V}\cdot\nabla\chi\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &\leq C\|\mathbf{V}\|_{(L^2(Q_T))^d}\|\chi\|_{(L^2(Q_T))^d}.\end{aligned}$$

Plugging the previous estimates into equation (5.86), one gets

$$\left|\langle\partial_t J(u_\eta),\chi\rangle\right|\leq\kappa\left(\|\chi\|_{L^\infty(Q_T)}+\|\chi\|_{L^2\left(0,T;H^1_0(\Omega)\right)}\right).$$

One can conclude the proof of part (iii), using the embedding of the Sobolev space $W^{1,q}(\Omega)\subset H^1_0(\Omega)\cap L^\infty(\Omega)$ for $q>d$, and consequently, one has

$$L^\infty\left(0,T;W^{1,q}(\Omega)\right)\subset L^2\left(0,T;H^1(\Omega)\right)\cap L^\infty\left(0,T;L^2(\Omega)\right),\quad\forall q>d.$$

To complete the proof of the lemma, we remark that the sequence $\left(J\left(u_\eta\right)\right)_\eta$ is lying into the Sobolev space

$$\mathcal{W}=\left\{J\left(u_\eta\right);J\left(u_\eta\right)\in L^2\left(0,T;H_0^1\left(\Omega\right)\right)\text{ and }\partial_t J\left(u_\eta\right)\in L^\infty\left(0,T;W^{-1,q'}\left(\Omega\right)\right)\right\}.$$

Thanks to the Aubin–Simon theorem, $\mathcal{W}$ is compactly embedded in $L^2\left(Q_T\right)$, and the sequence $\left(J\left(u_\eta\right)\right)_\eta$ is relatively compact in $L^2\left(0,T;L^2\left(\Omega\right)\right)$.

Since the differentiable function J is nondecreasing, then J^{-1} exists and it is continuous, then the sequence $\left(u_\eta\right)_\eta$ is relatively compact in $L^2\left(0,T;L^2\left(\Omega\right)\right)$. The proof of lemma 5.16 is now accomplished. □

Lemma 5.17. *Let $q_1=3r_1+2$ and $q_2=3r_2+2$, where r_1 and r_2 are given in assumption* (H4c). *The sequences $\left(1_{\{u_\eta\le u_*\}}u_\eta^{q_1}a\left(u_\eta\right)\nabla u_\eta\right)_\eta$ and $\left(1_{\{u_\eta\ge u_*\}}(1-u_\eta)^{q_2}a\left(u_\eta\right)\nabla u_\eta\right)_\eta$ are two Cauchy sequences in measure.*

Proof. The proof follows the same guidelines as the proof of lemma 5.14. We omit it for the sake of brevity. □

To conclude the proof of theorem 5.6, we deduce from lemma 5.16 and lemma 5.17, that we can extract a subsequence such that we have the following convergences

$$\begin{aligned}
&0\le u\left(\mathbf{x},t\right)\le 1\text{ a.e. in }Q_T,\\
&u_\eta\longrightarrow u\text{ strongly in }L^2\left(Q_T\right)\text{ and a.e. in }Q_T,\\
&J\left(u_\eta\right)\longrightarrow J\left(u\right)\text{strongly }L^2\left(Q_T\right),\\
&J\left(u_\eta\right)\longrightarrow J\left(u\right)\text{ weakly in }L^2\left(0,T;H_0^1\left(\Omega\right)\right),\\
&\sqrt{a\left(u_\eta\right)\mu'\left(u_\eta\right)}\nabla u_\eta\longrightarrow\sqrt{a\left(u\right)\mu'\left(u\right)}\nabla u\text{ weakly in }\left(L^2\left(Q_T\right)\right)^d,\\
&1_{\{u_\eta\le u_*\}}u_\eta^{q_1}a\left(u_\eta\right)\nabla u_\eta\longrightarrow 1_{\{u_\eta\le u_*\}}u^{q_1}a\left(u\right)\nabla u\text{ a.e. in }Q_T,\\
&1_{\{u_\eta\ge u_*\}}\left(1-u_\eta\right)^{q_2}a\left(u_\eta\right)\nabla u_\eta\longrightarrow 1_{\{u_\eta\ge u_*\}}\left(1-u\right)^{q_2}a\left(u\right)\nabla u\text{ a.e. in }Q_T.
\end{aligned}\tag{5.87}$$

We consider the following weak formulation

$$\begin{aligned}
&-\int_{Q_T}J_{\theta,\lambda}\left(u_\eta\right)\partial_t\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t-\int_\Omega J_{\theta,\lambda}\left(u_0\left(\mathbf{x}\right)\right)\chi\left(\mathbf{x},0\right)\mathrm{d}\mathbf{x}\\
&+\int_{Q_T}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t+\int_{Q_T}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla\chi j_{\theta,\lambda}\left(u_\eta\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&+\eta\int_{Q_T}\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t+\eta\int_{Q_T}\nabla u_\eta\cdot\nabla\chi j_{\theta,\lambda}\left(u_\eta\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&-\int_{Q_T}\left(f\left(u_\eta\right)-g\left(u_\eta\right)\right)\mathbf{V}\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t+\int_{Q_T}g'\left(u_\eta\right)\mathbf{V}\cdot\nabla u_\eta j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&-\int_{Q_T}\left(f\left(u_\eta\right)-g\left(u_\eta\right)\right)\mathbf{V}\cdot\nabla\chi j_{\theta,\lambda}\left(u_\eta\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t+\int_{Q_T}a\left(u_\eta\right)\mathbf{V}\cdot\nabla u_\eta j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&+\eta\int_{Q_T}\nabla u_\eta\cdot\mathbf{V}j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t=0,\quad\forall\chi\in\mathcal{C}^1\left([0,T];H_0^1\left(\Omega\right)\right)\text{ with }\chi\left(T,\cdot\right)=0
\end{aligned}\tag{5.88}$$

By splitting these integrals into two sub integrals, then we have the same convergence results obtained within the proof of theorem 5.4 for the integrals of the form $\displaystyle\int_{Q_T\cap\{u_\eta\le u_*\}}$ in (5.88).

Now, we show the convergence for the remaining terms, we have

$$\int_{\{u_\eta\ge u_*\}}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla\chi\,j_{\theta,\lambda}\left(u_\eta\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t\underset{\eta\to0}{\longrightarrow}\int_{\{u\ge u_*\}}a\left(u\right)\nabla u\cdot\nabla\chi\,j_{\theta,\lambda}\left(u\right)\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

Indeed, on the one hand, the sequence $\left(a\left(u_\eta\right)\nabla u_\eta\right)_\eta$ converges weakly in $\left(L^2\left(Q_T\right)\right)^d$ towards $a\left(u\right)\nabla u$. On the other hand, the sequence $\left(\nabla\chi\, j_{\theta,\lambda}\left(u_\eta\right)\right)_\eta$ converges strongly in $\left(L^2\left(Q_T\right)\right)^d$ towards $\nabla\chi\, j_{\theta,\lambda}\left(u\right)$. Furthermore, one obtains that

$$\eta\int_{\{u_\eta\geq u_*\}}\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\xrightarrow[\eta\to 0]{}0. \tag{5.89}$$

Indeed,

$$\eta\left|\int_{\{u_\eta\geq u_*\}}\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right|=\kappa\eta\int_{Q_T}\left|\left(1-u_\eta\right)^{\frac{r'}{2}-2+\lambda}\chi\,\nabla u_\eta\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t$$
$$\leq\kappa\eta\int_{Q_T}\left|\left(1-u_\eta\right)^{2r_2}\chi\nabla u_\eta\cdot\nabla u_\eta\right|\mathrm{d}\mathbf{x}\,\mathrm{d}t\leq\kappa\eta\|a\left(u_\eta\right)\nabla u_\eta\|_{\left(L^2(Q_T)\right)^d}\|\chi\|_{L^\infty(Q_T)}.$$

Then the convergence result (5.89) holds.
Now, we are interested in the third term of equation (5.88). We have $\left(a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\right)_\eta$ is a nonnegative sequence and

$$1_{\{u_\eta\geq u_*\}}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)=1_{\{u_\eta\geq u_*\}}c\left(u_*\right)\left(1-u\right)_\eta^{\frac{r'}{2}-2+\lambda}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla u_\eta$$

which converges almost everywhere, up to a subsequence, to $1_{\{u\geq u_*\}}a\left(u\right)\nabla u\cdot\nabla j_{\theta,\lambda}\left(u\right)$, since $\frac{r'}{2}-2+\lambda-2q_2-r_2\geq 0$, i.e. $\lambda\geq 7r_2+6-\frac{r'}{2}$.

Consider a nonnegative test function $\left(\chi\geq 0\right)$; then the Fatou's lemma ensures that

$$\liminf_{\eta\to 0}\int_{Q_T}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\geq\int_{Q_T}a\left(u\right)\nabla u\cdot\nabla j_{\theta,\lambda}\left(u\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

then the limit solution u verifies inequality (5.12) into definition 5.5. Finally, to obtain (5.13), we apply the Egorov theorem on the sequence $\left(a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\right)_\eta$ which converges almost everywhere. Indeed, we have

$$\begin{aligned}&\forall\varepsilon>0,\exists\, Q^\varepsilon\subset Q_T\text{ tel que mes }\left(Q^\varepsilon\right)<\varepsilon,\text{ and}\\&a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\xrightarrow[\eta\to 0]{}a\left(u\right)\nabla u\cdot\nabla j_{\theta,\lambda}\left(u\right)\text{ uniformly in }Q_T\backslash Q^\varepsilon.\end{aligned}$$

Now, we take a nonnegative test function χ such that $\mathrm{supp}\chi\subset\left([0,T)\times\Omega\right)\backslash Q^\varepsilon$, then

$$\int_{Q_T\backslash Q^\varepsilon}a\left(u_\eta\right)\nabla u_\eta\cdot\nabla j_{\theta,\lambda}\left(u_\eta\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\xrightarrow[\eta\to 0]{}\int_{Q_T\backslash Q^\varepsilon}a\left(u\right)\nabla u\cdot\nabla j_{\theta,\lambda}\left(u\right)\chi\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

This ends the proof of theorem 5.6. □

Technical Lemmas

We give in this appendix some technical results stemming from finite element discretizations. Let Ω be an open bounded polygonal subset of $\mathbb{R}^2$, and $\mathcal{T}$ be a conforming triangular mesh of Ω. For all triangle $T \in \mathcal{T}$, we denote by $h_T = \text{diam}\,(T)$, and by $h = \sup_{T\in\mathcal{T}} h_T$.
We consider the set $\mathcal{V}$ of the vertices of the triangles, and $(\mathbf{x}_K)_{K\in\mathcal{V}}$ their coordinates in Ω, and the usual $\mathbb{P}_1$ finite element space $\mathcal{H}_\mathcal{T}$ defined by

$$\mathcal{H}_\mathcal{T} = \{\phi \in C^0\left(\overline{\Omega}\right);\, \phi_{|_T} \in \mathbb{P}_1\left(\mathbb{R}\right),\quad \forall T \in \mathcal{T}\}.$$

Let $u_\mathcal{T} \in \mathcal{H}_\mathcal{T}$, then consider the piecewise constant functions $\overline{u}_\mathcal{T}$, $\underline{u}_\mathcal{T} : \Omega \longrightarrow \mathbb{R}$ defined by

$$\begin{aligned}
\overline{u}_\mathcal{T}(\mathbf{x}) &= \overline{u}_T = \sup_{\mathbf{x}\in T} u_\mathcal{T}(\mathbf{x}), \quad \text{if } \mathbf{x} \in T \in \mathcal{T},\\
\underline{u}_\mathcal{T}(\mathbf{x}) &= \underline{u}_T = \inf_{\mathbf{x}\in T} u_\mathcal{T}(\mathbf{x}), \quad \text{if } \mathbf{x} \in T \in \mathcal{T}.
\end{aligned}$$

Lemma A.1. *There exists an absolute constant $c > 0$ such that*

$$\int_\Omega \left|\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})\right|^2 \mathrm{d}\mathbf{x} \le ch^2 \int_\Omega \left|\nabla u_\mathcal{T}(\mathbf{x})\right|^2 \mathrm{d}\mathbf{x},$$

where $c = \frac{243}{2\pi^2}$.

Proof. Let $T \in \mathcal{T}$ be a triangle whose vertices are located at $\mathbf{x}_K$, $\mathbf{x}_L$, and $\mathbf{x}_M$. We assume without loss of generality that

$$\underline{u}_T = u_M \le u_L \le u_K = \overline{u}_T \tag{A.1}$$

We denote by $\omega^T_{K,L}$ the triangle whose vertices are $\mathbf{x}_K$, $\mathbf{x}_{KL}$, and $\mathbf{x}_T$, where $\mathbf{x}_T$ and $\mathbf{x}_{KL}$ represent respectively, the center of gravity and the midpoint of $[\mathbf{x}_K, \mathbf{x}_L]$ of the triangle T. Specifically, we have

$$\mathbf{x}_T = \frac{1}{3}(\mathbf{x}_K + \mathbf{x}_L + \mathbf{x}_M), \quad \mathbf{x}_{KL} = \frac{1}{2}(\mathbf{x}_K + \mathbf{x}_L).$$

Note that

$$\left|\omega^T_{K,L}\right| = \frac{|T|}{6}, \tag{A.2}$$

where $\left|\omega_{K,L}^T\right|$ (respectively $|T|$) denotes the Lebesgue measure of $\omega_{K,L}^T$ (respectively T). Since $u_{\mathcal{T}} \in \mathcal{H}_{\mathcal{T}}$ is a linear function of $\mathbf{x}$ on T, then one has

$$u_T := \frac{1}{|T|}\int_T u_{\mathcal{T}}(\mathbf{x})\,\mathrm{d}\mathbf{x} = u(\mathbf{x}_T) = \frac{1}{3}(u_K + u_L + u_M),$$

and in view of property (A.1), one has

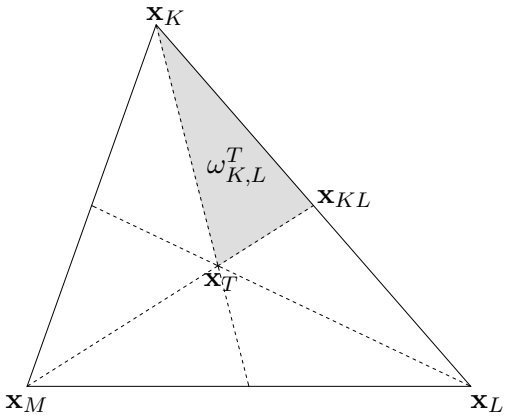


FIGURE A.1 – The triangle $T \in \mathcal{T}$ and the sub-triangle $w_{K,L}^T$.

$$u_{\mathcal{T}}(\mathbf{x}_{KL}) = \frac{1}{2}(u_K + u_L) \geq u_T.$$

As a consequence,

$$u_{\mathcal{T}}(\mathbf{x}) \geq u_T, \quad \forall \mathbf{x} \in \omega_{K,L}^T. \tag{A.3}$$

Therefore, one has

$$\begin{aligned}\int_T |u_{\mathcal{T}}(\mathbf{x}) - u_T|^2\,\mathrm{d}\mathbf{x} &\geq \int_{\omega_{K,L}^T} (u_{\mathcal{T}}(\mathbf{x}) - u_T)^2\,\mathrm{d}\mathbf{x} = \frac{\left|\omega_{K,L}^T\right|}{9}(u_K + u_{KL} - 2u_T)^2 \\ &= \frac{\left|\omega_{K,L}^T\right|}{9}\left(u_K + \frac{u_K + u_L}{2} - 2u_T\right)^2 = \frac{\left|\omega_{K,L}^T\right|}{324}(5u_K - u_L - 4u_M)^2 \\ &\geq \frac{2|T|}{243}(u_K - u_M)^2 = \frac{2}{243}\int_T |\overline{u}_{\mathcal{T}}(\mathbf{x}) - \underline{u}_{\mathcal{T}}(\mathbf{x})|^2\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{A.4}$$

On the other hand, it follows from the Poincaré-Wirtinger inequality that (see [70])

$$\int_T |u_{\mathcal{T}}(\mathbf{x}) - u_T|^2\,\mathrm{d}\mathbf{x} \leq \frac{{h_T}^2}{\pi^2}\int_T |\nabla u_{\mathcal{T}}(\mathbf{x})|^2\,\mathrm{d}\mathbf{x},$$

which, together with (A.4), yields

$$\begin{aligned}\int_\Omega |\overline{u}_{\mathcal{T}}(\mathbf{x}) - \underline{u}_{\mathcal{T}}(\mathbf{x})|^2\,\mathrm{d}\mathbf{x} &= \sum_{T\in\mathcal{T}}\int_T |\overline{u}_{\mathcal{T}}(\mathbf{x}) - \underline{u}_{\mathcal{T}}(\mathbf{x})|^2\,\mathrm{d}\mathbf{x} \\ &\leq \frac{243}{2}\sum_{T\in\mathcal{T}}\int_T |u_{\mathcal{T}}(\mathbf{x}) - u_T|^2\,\mathrm{d}\mathbf{x} \leq ch^2\int_\Omega |\nabla u_{\mathcal{T}}(\mathbf{x})|^2\,\mathrm{d}\mathbf{x}.\end{aligned}$$

This ends the proof of the lemma. □

Remark 5. We can easily see that Lemma A.1 holds in L^1. Indeed, using inequality (A.3) and the Poincaré-Wirtinger inequality in L^1 for convex domains (see [1]), one gets

$$\int_\Omega |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})| \, \mathrm{d}\mathbf{x} \leq \frac{27h}{2} \int_\Omega |\nabla u_\mathcal{T}(\mathbf{x})| \, \mathrm{d}\mathbf{x}.$$

Lemma A.2. *Let $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$, and let $u_\mathcal{T}$ and $u_\mathcal{M}$ be respectively the corresponding piecewise linear and piecewise constant reconstructions, then*

$$\int_\Omega |u_\mathcal{T}(\mathbf{x}) - u_\mathcal{M}(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \leq ch^2 \|\nabla u_\mathcal{T}\|^2_{(L^2(\Omega))^2}, \tag{A.5}$$

where c is the same constant given in Lemma A.1.

Proof. In order to prove inequality (A.5), we first prove the following inequality

$$|u_\mathcal{T}(\mathbf{x}) - u_\mathcal{M}(\mathbf{x})| \leq |\overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x})|, \qquad \text{for a.e. } \mathbf{x} \in \Omega. \tag{A.6}$$

Let $\mathbf{x} \in \Omega = \cup_{T\in\mathcal{T}} \overline{T}$, there exists a triangle T such that $\mathbf{x} \in \overline{T}$. We assume, without loss of generality, that $\mathbf{x}$ lies in the interior of the triangle $\omega^T_{K,L}$.
One has, from the definition of $\overline{u}_\mathcal{T}$ and $\underline{u}_\mathcal{T}$, that

$$u_\mathcal{T}(\mathbf{x}) - u_\mathcal{M}(\mathbf{x}) = u_\mathcal{T}(\mathbf{x}) - u_K \leq \overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x}).$$

On the other hand, we have $u_\mathcal{M}(\mathbf{x}) = u_K \leq \overline{u}_\mathcal{T}(\mathbf{x})$ and since $\underline{u}_\mathcal{T}(\mathbf{x}) \leq u_\mathcal{T}(\mathbf{x})$, one gets

$$u_\mathcal{M}(\mathbf{x}) - u_\mathcal{T}(\mathbf{x}) \leq \overline{u}_\mathcal{T}(\mathbf{x}) - \underline{u}_\mathcal{T}(\mathbf{x}).$$

One can conclude the proof of this lemma by using Lemma A.1 as well as inequality (A.6). □

Let $T \in \mathcal{T}$ be a triangle whose vertices are located at $\mathbf{x}_K$, $\mathbf{x}_L$, and $\mathbf{x}_M$. We denote by $\mathbf{x}_T$ its center of gravity, by $\mathbf{x}_{ML}$ the midpoint of $[\mathbf{x}_M, \mathbf{x}_L]$, and by $\mathbf{x}_{MK}$ the midpoint of $[\mathbf{x}_M, \mathbf{x}_K]$. We denote by ω_M the dual control volume constructed around $\mathbf{x}_M$ and by ω^T_M the intersection between ω_M and T.

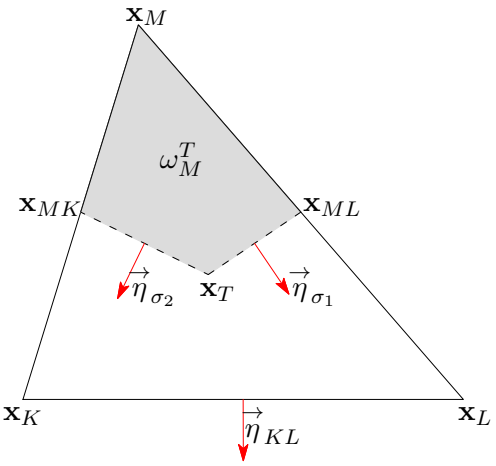


FIGURE A.2 – The triangle $T \in \mathcal{T}$ and the normal vectors on the edges.

Lemma A.3. *With the notations of figure A.2, the following geometric property*

$$\sum_{i=1}^{2} |\sigma_i| \overrightarrow{\eta}_{\sigma_i} = \frac{1}{2} |\mathbf{x}_K \mathbf{x}_L| \overrightarrow{\eta}_{KL}$$

holds.

Proof. Applying the Green–Gauss formula to the constant field vector $\overrightarrow{1}$ over the domain ω_M^T, one has

$$\int_{\omega_M^T} \operatorname{div}\left(\overrightarrow{1}\right) \mathrm{d}\mathbf{x} = 0 = \int_{\partial\omega_M^T} \overrightarrow{\eta}\, \mathrm{d}\sigma = \sum_{\sigma \subset \partial\omega_M^T} |\sigma| \overrightarrow{\eta}_\sigma = \sum_{i=1}^{4} |\sigma_i| \overrightarrow{\eta}_{\sigma_i}, \tag{A.7}$$

where σ_3 (resp. σ_4) represents the length of $[\mathbf{x}_M, \mathbf{x}_{ML}]$ (resp. $[\mathbf{x}_M, \mathbf{x}_{MK}]$) and $\overrightarrow{\eta}_{\sigma_i}$ represents the unit normal vector to $\sigma_i, i = 3, 4$ outward to T.
On the other hand, one applies again the Green–Gauss formula over the triangle T and gets

$$\begin{aligned} \int_T \operatorname{div}\left(\overrightarrow{1}\right) \mathrm{d}\mathbf{x} = 0 = \int_{\partial T} \overrightarrow{\eta}\, \mathrm{d}\sigma &= \sum_{\sigma \subset \partial T} |\sigma| \overrightarrow{\eta}_\sigma \\ &= |\mathbf{x}_M \mathbf{x}_L| \overrightarrow{\eta}_{\sigma_3} + |\mathbf{x}_M \mathbf{x}_K| \overrightarrow{\eta}_{\sigma_4} + |\mathbf{x}_K \mathbf{x}_L| \overrightarrow{\eta}_{KL}. \end{aligned} \tag{A.8}$$

One can conclude the proof of the lemma by combining equations (A.7)–(A.8). □

A.1 The reference element

We recall that in *the control volume finite element method*, we have introduced the transmissibility coefficients

$$\Lambda_{KL} = \int_\Omega \Lambda(\mathbf{x}) \nabla\varphi_K(\mathbf{x}) \cdot \nabla\varphi_L(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

In a computation point of view, and in order to approximate these elements, we need : either (a) an effective way of evaluating the functions $(\varphi_K)_{K\in\mathcal{V}}$ and their gradients, or (b) a closed form for the resulting integrals. Both possibilities are done usually by introducing the so-called reference element.
For triangles, the reference element $\widehat{T}$ is the triangle with vertices

$$\widehat{S_1} = (0,0)\,, \quad \widehat{S_2} = (1,0)\,, \quad \widehat{S_1} = (0,1)\,.$$

To distinguish variables in the reference element and in a physical triangle (a general triangle), we usually adopt the variables $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ in the reference element and $(\mathbf{x}, \mathbf{y})$ in the physical element. We designate by T a physical triangle with vertices $S_1 = (\mathbf{x}_1, \mathbf{y}_1)$, $S_2 = (\mathbf{x}_2, \mathbf{y}_2)$, and $S_3 = (\mathbf{x}_3, \mathbf{y}_3)$. We denote by $(\mathbf{x}_{ji}, \mathbf{y}_{ji}) = (\mathbf{x}_j - \mathbf{x}_i, \mathbf{y}_j - \mathbf{y}_i)$ the coordinates of the vector $\overrightarrow{S_i S_j}$, for every $i \neq j \in \{1,2,3\}$.

Let $\mathcal{F}_T$ be the affine transformation mapping the triangle $\widehat{T}$ bijectively into T. We have $\mathcal{F}_T\left(\widehat{S}_i\right) = S_i$ for every $i = 1, 2, 3$. Thus we have

$$\begin{aligned} \mathcal{F}_T \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} &= \begin{pmatrix} \mathbf{x}_{21} & \mathbf{x}_{31} \\ \mathbf{y}_{21} & \mathbf{y}_{31} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{pmatrix} + \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix} (1 - \hat{\mathbf{x}} - \hat{\mathbf{y}}) + \begin{pmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{pmatrix} \hat{\mathbf{x}} + \begin{pmatrix} \mathbf{x}_3 \\ \mathbf{y}_3 \end{pmatrix} \hat{\mathbf{y}}. \end{aligned} \tag{A.9}$$

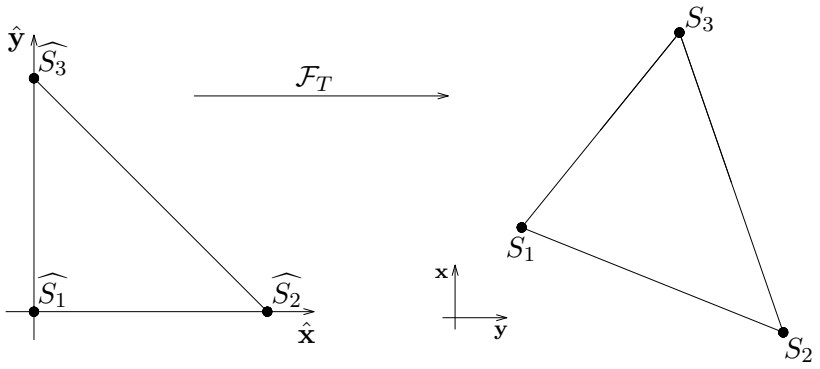


FIGURE A.3 – Reference and physical triangles.

The local nodal functions in the reference triangles are three $\mathbb{P}_1$ functions satisfying

$$\widehat{\varphi}_i\left(\widehat{S_j}\right) = \delta_{ij}, \qquad i,j = 1,2,3.$$

They are precisely defined, into the space of coordinates $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, by

$$\begin{aligned}
&\widehat{\varphi}_1(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = 1 - \hat{\mathbf{x}} - \hat{\mathbf{y}},\\
&\widehat{\varphi}_2(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \hat{\mathbf{x}},\\
&\widehat{\varphi}_3(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \hat{\mathbf{y}}.
\end{aligned}$$

It is simple now to prove that

$$\widehat{\varphi}_i = \varphi_i \circ \mathcal{F}_T, \qquad i = 1,2,3.$$

or, what is the same

$$\varphi_i = \widehat{\varphi}_i \circ \mathcal{F}_T^{-1}, \qquad i = 1,2,3. \tag{A.10}$$

We denote by $\mathbf{B}_T$ the matrix of the linear transformation $\mathcal{F}_T$, we have

$$\det(\mathbf{B}_T) = 2|T| \neq 0 \quad \text{and} \quad \mathcal{F}_T^{-1}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{x}}\\ \hat{\mathbf{y}}\end{pmatrix} = \mathbf{B}_T^{-1}\begin{pmatrix}\mathbf{x}-\mathbf{x}_1\\ \mathbf{y}-\mathbf{y}_1\end{pmatrix}, \tag{A.11}$$

where $\mathbf{B}_T^{-1} = \dfrac{1}{2|T|}\begin{pmatrix}\mathbf{y}_{31} & -\mathbf{x}_{31}\\ -\mathbf{y}_{21} & \mathbf{x}_{21}\end{pmatrix}$ is the inverse of the matrix $\mathbf{B}_T$.

Using together equations (A.10) and (A.11), one has a simple way of evaluating the nodal functions $(\varphi_i)_{i=1,2,3}$ and therefore their gradients. For instance, one gets

$$\begin{aligned}
\varphi_1(\mathbf{x},\mathbf{y}) &= 1 - \frac{1}{2|T|}\left(\mathbf{y}_{32}(\mathbf{x}-\mathbf{x}_1) - \mathbf{x}_{32}(\mathbf{y}-\mathbf{y}_1)\right),\\
\varphi_2(\mathbf{x},\mathbf{y}) &= \frac{1}{2|T|}\left(\mathbf{y}_{31}(\mathbf{x}-\mathbf{x}_1) - \mathbf{x}_{31}(\mathbf{y}-\mathbf{y}_1)\right),\\
\varphi_3(\mathbf{x},\mathbf{y}) &= \frac{1}{2|T|}\left(-\mathbf{y}_{21}(\mathbf{x}-\mathbf{x}_1) + \mathbf{x}_{21}(\mathbf{y}-\mathbf{y}_1)\right),
\end{aligned}$$

and

$$\nabla\varphi_1(\mathbf{x},\mathbf{y}) = \frac{-1}{2|T|}\begin{pmatrix} \mathbf{y}_{32} \\ -\mathbf{x}_{32} \end{pmatrix} = \frac{-||\overrightarrow{S_2S_3}||}{2|T|}\overrightarrow{\eta}_{23},$$

$$\nabla\varphi_2(\mathbf{x},\mathbf{y}) = \frac{1}{2|T|}\begin{pmatrix} \mathbf{y}_{31} \\ -\mathbf{x}_{31} \end{pmatrix} = \frac{-||\overrightarrow{S_3S_1}||}{2|T|}\overrightarrow{\eta}_{31},$$

$$\nabla\varphi_3(\mathbf{x},\mathbf{y}) = \frac{-1}{2|T|}\begin{pmatrix} \mathbf{y}_{21} \\ -\mathbf{x}_{21} \end{pmatrix} = \frac{-||\overrightarrow{S_1S_2}||}{2|T|}\overrightarrow{\eta}_{12},$$

where $\overrightarrow{\eta}_{ij} = (-\mathbf{y}_{ij}, \mathbf{x}_{ij})$ is the unit normal vector at the line segment $[S_iS_j]$ **outward** to T, for every $i \neq j \in \{1,2,3\}$.

Bibliographie

[1] Gabriel Acosta and Ricardo G Durán. An optimal poincaré inequality in l 1 for convex domains. *Proceedings of the American Mathematical Society*, pages 195–202, 2004. 153

[2] Mohamed Afif and Brahim Amaziane. Convergence of finite volume schemes for a degenerate convection–diffusion equation arising in flow in porous media. *Computer methods in applied mechanics and engineering*, 191(46) :5265–5286, 2002. 54, 55

[3] Abdellatif Agouzal, Jacques Baranger, Jean-FranÃğois Maitre, and Fabienne Oudin. Connection between finite volume and mixed finite element methods for a diffusion problem with nonconstant coefficients. application to a convection diffusion problem. *EAST WEST J NUMER MATH*, 3(4) :237–254, 1995. 54

[4] Herbert Amann. Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems. *Function spaces, differential operators and nonlinear analysis*, 133 :9–126, 1993. 32

[5] Youcef Amirat, Kamel Hamdache, and Abdelhamid Ziani. Mathematical analysis for compressible miscible displacement models in porous media. *Mathematical Models and Methods in Applied Sciences*, 6(06) :729–747, 1996. 120

[6] Youcef Amirat and Abdelhamid Ziani. Global weak solutions for a parabolic system modeling a one-dimensional miscible flow in porous media. *Journal of mathematical analysis and applications*, 220(2) :697–718, 1998. 120

[7] Boris Andreianov, Mostafa Bendahmane, and Mazen Saad. Finite volume methods for degenerate chemotaxis model. *Journal of computational and applied mathematics*, 235(14) :4015–4031, 2011. 31, 47, 54, 55

[8] Mostafa Bendahmane, Kenneth H Karlsen, and José Miguel Urbano. On a two-sidedly degenerate chemotaxis model with volume-filling effect. *Mathematical Models and Methods in Applied Sciences*, 17(05) :783–804, 2007. 4, 31, 120

[9] Mostafa Bendahmane and Mazen Saad. Mathematical analysis and pattern formation for a partial immune system modeling the spread of an epidemic disease. *Acta applicandae mathematicae*, 115(1) :17–42, 2011. 30, 54

[10] Fred Brauer and Carlos Castillo-Chavez. *Mathematical models in population biology and epidemiology*. Springer, 2011. 56

[11] Konstantin Brenner, Roland Masson, et al. Convergence of a vertex centred discretization of two-phase darcy flows on general meshes. *International Journal of Finite Volume*, 10 :1–37, 2013. 96

[12] Haïm Brezis. *Analyse fonctionelle :Théorie et applications*, volume 5. Masson, 1983. 74

[13] Giuseppe Buttazzo, Gérard Michaille, and Hedy Attouch. *Variational analysis in Sobolev and BV spaces : applications to PDEs and optimization*, volume 6. Siam, 2006. 74

[14] HM Byrne and MAJ Chaplain. Mathematical models for tumour angiogenesis : numerical simulations and nonlinear wave solutions. *Bulletin of Mathematical Biology*, 57(3) :461–486, 1995. 30, 54

[15] Zhiqiang Cai. On the finite volume element method. *Numerische Mathematik*, 58(1) :713–735, 1990. 54

[16] Clément Cancès and Cindy Guichard. Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. HAL : hal-00955091, 2014. 73, 86, 94, 103, 104, 105

[17] Emilio Cariaga, Fernando Concha, Iuliu Sorin Pop, and Mauricio Sepúlveda. Convergence analysis of a vertex-centered finite volume scheme for a copper heap leaching model. *Mathematical Methods in the Applied Sciences*, 33(9) :1059–1077, 2010. 54, 55

[18] Juan Casado-Díaz, T Chacón Rebollo, Vivette Girault, M Gómez Mármol, and François Murat. Finite elements approximation of second order linear elliptic equations in divergence form with right-hand side in l 1. *Numerische Mathematik*, 105(3) :337–374, 2007. 63

[19] Mark AJ Chaplain, Mahadevan Ganesh, and Ivan G Graham. Spatio-temporal pattern formation on spherical surfaces : numerical simulation and application to solid tumour growth. *Journal of mathematical biology*, 42(5) :387–423, 2001. 30, 54

[20] Guy Chavent, Jérôme Jaffré, and Jean E Roberts. Mixed-hybrid finite elements and cell-centred finite volumes for two-phase flow in porous media. *Mathematical Modelling of Flow Through Porous Media*, pages 100–114, 1995. 54

[21] John Condeelis, Robert H Singer, and Jeffrey E Segall. The great escape : when cancer cells hijack the genes for chemotaxis and motility. *Annu. Rev. Cell Dev. Biol.*, 21 :695–718, 2005. 4

[22] Yves Coudière, Jean-Paul Vila, and Philippe Villedieu. Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. *ESAIM : Mathematical Modelling and Numerical Analysis*, 33(03) :493–516, 1999. 54

[23] Klaus Deimling. Nonlinear functional analysis. 1985. *Springler-Verlag, Berlin*. 105

[24] Dirk Dormann and Cornelis J Weijer. Chemotactic cell movement during development. *Current opinion in genetics & development*, 13(4) :358–364, 2003. 4

[25] Jérôme Droniou, Robert Eymard, Thierry Gallouët, and Raphaele Herbin. Gradient schemes : a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. *Mathematical Models and Methods in Applied Sciences*, 23(13) :2395–2432, 2013. 54

[26] Renjun Duan, Alexander Lorz, and Peter Markowich. Global solutions to the coupled chemotaxis-fluid equations. *Communications in Partial Differential Equations*, 35(9) :1635–1673, 2010. 120

[27] Alexandre Ern and Jean-Luc Guermond. Theory and practice of finite elements, vol. 159 of applied mathematical sciences, 2004. 96

[28] Lawrence C Evans. Partial differential equations : Graduate studies in mathematics. *American Mathematical Society*, 2, 1998. 69, 125

[29] Robert Eymard and Thierry Gallouët. Convergence d'un schéma de type éléments finis-volumes finis pour un système formé d'une équation elliptique et d'une équation hyperbolique. *Modélisation mathématique et analyse numérique*, 27(7) :843–861, 1993. 54

[30] Robert Eymard, Thierry Gallouët, Mustapha Ghilani, and Raphaèle Herbin. Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. *IMA J. Numer. Anal.*, 18(4) :563–594, 1998. 105

[31] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. *Handbook of numerical analysis*, 7 :713–1018, 2000. 7, 41, 46, 54, 75, 91

[32] Robert Eymard, Thierry Gallouët, and Raphaele Herbin. A finite volume scheme for anisotropic diffusion problems. *Comptes Rendus Mathématique*, 339(4) :299–302, 2004. 54, 92, 109

[33] Robert Eymard, Thierry Gallouët, and Raphaele Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes sushi : a scheme using stabilization and hybrid interfaces. *IMA Journal of Numerical Analysis*, 30(4) :1009–1043, 2010. 54

[34] Robert Eymard, Thierry Gallouët, Raphaèle Herbin, and Anthony Michel. Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numerische Mathematik*, 92(1) :41–82, 2002. 110

[35] Pierre Fabrie and Michel Langlais. Mathematical analysis of miscible displacement in porous medium. *SIAM journal on mathematical analysis*, 23(6) :1375–1392, 1992. 120

[36] Francis Filbet. A finite volume scheme for the patlak–keller–segel chemotaxis model. *Numerische Mathematik*, 104(4) :457–488, 2006. 47

[37] Peter A Forsyth. A control volume finite element approach to napl groundwater contamination. *SIAM Journal on Scientific and Statistical Computing*, 12(5) :1029–1057, 1991. 55

[38] Thierry Gallouët and Jean-Claude Latché. Compactness of discrete approximate solutions to parabolic pdes-application to a turbulence model. *Communications on Pure & Applied Analysis*, 11(6), 2012. 109

[39] Cédric Galusinski and Mazen Saad. On a degenerate parabolic system for compressible, immiscible, two-phase flows in porous media. *Advances in Differential Equations*, 9(11-12) :1235–1278, 2004. 120

[40] Sergueï Godounov, A Zabrodine, Mikhail Ivanov, A Kraiko, and G Prokopov. *Résolution numérique des problemes multidimensionnels de la dynamique des gaz*. Editions Mir, 1979. 92

[41] SK Godunov, A Zabrodine, M Ivanov, A Kraiko, and G Prokopov. Résolution numérique des problèmes multidimensionnels de la dynamique des gaz. 1979. *Mir, Moscow*. 91

[42] Raphaèle Herbin and Florence Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. *Finite volumes for complex applications V*, pages 659–692, 2008. 55

[43] E Hildebrand and UB Kaupp. Sperm chemotaxis : a primer. *Annals of the New York Academy of Sciences*, 1061(1) :221–225, 2005. 4

[44] Thomas Hillen and Kevin Painter. Global existence for a parabolic chemotaxis model with prevention of overcrowding. *Advances in Applied Mathematics*, 26(4) :280–301, 2001. 5

[45] Thomas Hillen and Kevin J Painter. A userâĂŹs guide to pde models for chemotaxis. *Journal of mathematical biology*, 58(1-2) :183–217, 2009. 56

[46] Dirk Horstmann. From 1970 until present : the keller-segel model in chemotaxis and its consequences. i. *Jahresbericht der Deutschen Mathematiker Vereinigung*, 105 :103–165, 2003. 31

[47] Dirk Horstmann. From 1970 until present : The keller-segel model in chemotaxis and its consequences ii. *Jahresbericht der Deutschen Mathematiker Vereinigung*, 106(2) :51–70, 2004. 4, 30, 31, 54

[48] Hyung Ju Hwang, Kyungkeun Kang, and Angela Stevens. Drift-diffusion limits of kinetic models for chemotaxis : a generalization. *Discrete Contin. Dyn. Syst. Ser. B*, 5(2) :319–334, 2005. 37

[49] Moustafa Ibrahim and Mazen Saad. Pattern formation and cross-diffusion for a chemotaxis model. In *Conference Book Biomath 2013*, 2013. 79

[50] Moustafa Ibrahim and Mazen Saad. On the efficacy of a control volume finite element method for the capture of patterns for a volume-filling chemotaxis model. *Computers & Mathematics with Applications*, 2014. 86, 93

[51] Evelyn F Keller and Lee A Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3) :399–415, 1970. 4, 30, 54

[52] Evelyn F Keller and Lee A Segel. Model for chemotaxis. *Journal of Theoretical Biology*, 30(2) :225–234, 1971. 30, 54

[53] Bruno Larrivee and Aly Karsan. Signaling pathways induced by vascular endothelial growth factor (review). *International journal of molecular medicine*, 5(5) :447–503, 2000. 4

[54] Jean Leray and Juliusz Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup.(3)*, 51 :45–78, 1934. 105

[55] Jacques-Louis Lions, Jacques-Louis Lions, Jacques-Louis Lions, and Jacques-Louis Lions. *Quelques méthodes de résolution des problemes aux limites non linéaires*, volume 31. Dunod Paris, 1969. 69

[56] Pierre-Louis Lions. Mathematical topics in fluid mechanics : Volume 1 : Incompressible models. 1996. 125

[57] Alexander Lorz. Coupled chemotaxis fluid model. *Mathematical Models and Methods in Applied Sciences*, 20(06) :987–1004, 2010. 120

[58] Philip K Maini. Using mathematical models to help understand biological pattern formation. *Comptes rendus biologies*, 327(3) :225–234, 2004. 30

[59] James D Murray. How the leopard gets its spots. *Scientific American*, 258(3) :80–87, 1988. 2, 54

[60] James D Murray. *Mathematical biology : I. An introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer, 2002. 30, 33

[61] James D Murray. *Mathematical biology II : spatial models and biomedical applications*, volume 18 of *Interdisciplinary Applied Mathematics*. Springer, New York, 2003. 2, 3, 30, 33, 34

[62] James D Murray. Turing centennial celebration. Princeton University, May 10-12 2012. 2

[63] James D Murray and MR Myerscough. Pigmentation pattern formation on snakes. *Journal of theoretical biology*, 149(3) :339–360, 1991. 30, 54

[64] Mario Ohlberger. A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations. *ESAIM : Mathematical Modelling and Numerical Analysis*, 35(02) :355–387, 2001. 55

[65] Stanley Osher and Fred Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Mathematics of computation*, 38(158) :339–374, 1982. 92

[66] Kevin J Painter and Thomas Hillen. Volume-filling and quorum-sensing in models for chemosensitive movement. *Can. Appl. Math. Quart*, 10(4) :501–543, 2002. 31, 56

[67] KJ Painter, PK Maini, and HG Othmer. Stripe formation in juvenile pomacanthus explained by a generalized turing mechanism with chemotaxis. *Proceedings of the National Academy of Sciences*, 96(10) :5549–5554, 1999. 30, 54

[68] Hwan Tae Park, Jane Wu, and Yi Rao. Molecular control of neuronal migration. *Bioessays*, 24(9) :821–827, 2002. 4

[69] Clifford S Patlak. Random walk with persistence and external bias. *The Bulletin of mathematical biophysics*, 15(3) :311–338, 1953. 4, 30, 54

[70] Lawrence E Payne and Hans F Weinberger. An optimal poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1) :286–292, 1960. 152

[71] Alex Potapov and Thomas Hillen. Metastability in chemotaxis models. *Journal of Dynamics and Differential Equations*, 17(2) :293–330, 2005. 31

[72] P Schaaf and J Talbot. Surface exclusion effects in adsorption processes. *The Journal of chemical physics*, 91 :4401, 1989. 31

[73] Jacques Simon. Compact sets in the space $L^p(0,T;B)$. *Annali di Matematica pura ed applicata*, 146(1) :65–96, 1986. 124

[74] Angela Stevens and Hans G Othmer. Aggregation, blowup, and collapse : The abc's of taxis in reinforced random walks. *SIAM Journal on Applied Mathematics*, 57(4) :1044–1081, 1997. 56

[75] Roger Temam. Navier-stokes equations. theory and numerical analysis. reprint of the 1984 edition. *AMS Chelsea, Providence, RI*, 2001. 69

[76] Alan Mathison Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641) :37–72, 1952. 2, 30, 54

[77] R Tyson, SR Lubkin, and James D Murray. A minimal mechanism for bacterial pattern formation. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 266(1416) :299–304, 1999. 53

[78] Zhian Wang and Thomas Hillen. Classical solutions and pattern formation for a volume filling chemotaxis model. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 17(3) :037108–037108, 2007. 5, 7, 31, 32, 56, 79

[79] Dariusz Wrzosek. Global attractor for a chemotaxis model with prevention of overcrowding. *Nonlinear Analysis : Theory, Methods & Applications*, 59(8) :1293–1310, 2004. 5

[80] Xuesong Yang, Dirk Dormann, Andrea E Münsterberg, and Cornelis J Weijer. Cell movement patterns during gastrulation in the chick are controlled by positive and negative chemotaxis mediated by fgf4 and fgf8. *Developmental cell*, 3(3) :425–437, 2002. 30

[81] Anatol M Zhabotinsky. Periodical oxidation of malonic acid in solution (a study of the belousov reaction kinetics). *Biofizika*, 9 :306–311, 1964. 29

l'université nantes angers le mans
COMMUNAUTÉ D'UNIVERSITÉS ET ÉTABLISSEMENTS

Centrale Nantes

Thèse de Doctorat

Moustafa IBRAHIM

Systèmes paraboliques dégénérés intervenant en mécanique des fluides et en médecine: analyse mathématique et numérique

Degenerate parabolic systems involved in fluid mechanics and medicine: mathematical and numerical analysis

Résumé

Dans cette thèse, nous nous intéressons à l'analyse mathématique et numérique des systèmes paraboliques non linéaires dégénérés découlant, soit de la modélisation de la chimiotaxie, soit de la modélisation des fluides compressibles. Le modèle de chimiotaxie (Keller-Segel) proposé est un modèle de dynamique des populations décrivant l'évolution spatio-temporelle de la densité cellulaire et de la concentration chimiotactique. Pour ce modèle, nous étudions la formation de patterns en utilisant l'analyse de stabilité linéaire et le principe de Turing. Nous proposons ensuite un schéma numérique CVFE pour un modèle anisotrope de Keller-Segel. La construction de ce schéma est basée sur la méthode des éléments finis pour le terme de diffusion et sur la méthode des volumes finis classique pour le terme de convection. Nous montrons que ce schéma assure le principe de maximum discret et qu'il est consistent dans le cas où tous les coefficients de transmissibilité sont positifs. Par la suite, sur des maillages triangulaires généraux, nous proposons et analysons un schéma numérique CVFE non linéaire. Ce schéma est basé sur l'utilisation d'un flux numérique de Godunov pour le terme de diffusion, tandis que le terme de convection est approché au moyen d'un décentrage amont et d'un flux de Godunov. D'une part, le décentrage amont permet d'avoir le principe de maximum. D'autre part, le flux de Godunov assure que les solutions discrètes soient bornées sans restriction sur le maillage du domaine spatial ni sur les coefficients de transmissibilité. Nous réalisons différentes simulations numériques bi-dimensionnelles pour illustrer l'efficacité du schéma à tenir compte des hétérogénéités. Enfin, nous nous intéressons à une équation parabolique dégénérée contenant des termes dégénérés d'ordre 0 et 1 et décrivant un modèle de chimiotaxie-fluide ou l'écoulement d'un fluide compressible. Une formulation faible classique est souvent possible en absence des termes dégénérés d'ordre 0 et 1 ; tandis que dans le cas général, nous obtenons des solutions dans un sens affaibli vérifiant une formulation de type inégalité variationnelle. La définition des solutions faibles est adaptée à la nature de la dégénérescence des termes de dissipation.

Mots clés

Systèmes de réaction-diffusion, Formation de patterns, Chimiotaxie, Stabilité linéaire, Méthode de volumes finis, Méthode des éléments finis, Systèmes paraboliques dégénérés, Tenseurs anisotropes hétérogènes, Fluide compressible.

Abstract

In this thesis, we are interested in the mathematical and numerical analysis of nonlinear degenerate parabolic systems arising either from modeling the chemotaxis process, or from modeling compressible flows in porous media. The proposed chemotaxis model (Keller-Segel model) is a model of population dynamics describing the spatio-temporel evolution of the cell density and the chemical concentration. For this model, we study the pattern formation using the linear stability analysis as well as the principle of Turing. Then, we propose a numerical scheme (CVFE scheme) for an anisotropic Keller-Segel model. The construction of the scheme is based on the use of each of the finite element scheme for the diffusion term and the upwind finite volume scheme for the convective term. We show that the scheme is consistent and ensures the discrete maximum principle in the case where all the transmissibility coefficients are nonnegative. Thereafter, over general triangular meshes, we propose and analyze a nonlinear CVFE scheme. This scheme is based on the use of the Godunov flux function for the diffusion term, while the convective term is approximated by parts using an upwind finite volume scheme and a Godunov flux function. First, the upwind finite volume scheme allows of having the discrete maximum principle. On the other hand, the Godunov scheme ensures the boundedness of the discrete solutions without restrictions on the mesh nor on the transmissibility coefficients. Using this scheme, we realize some numerical simulations to illustrate the effectiveness of the scheme. Finally, we are interested in a degenerate parabolic equation containing degenerate terms of order 0 and 1 and describing a chemotaxis-fluid model or a displacement of compressible flows. Classical weak formulation is often possible in the absence of degenerate terms of order 0 and 1; while in the general case, we obtain weak solutions in the sense of verifying a weighted formulation. The definition of weak solutions is adapted to the nature of the degeneracy of the dissipative terms.

Key Words

Reaction-diffusion systems, Pattern formation, Chemotaxis, Linear stability, Finite volume method, Finite element method, Degenerate parabolic systems, Heterogeneous anisotropic tensors, Compressible fluid.